Learning to Field Test in Policing:

Using an analysis of completed randomised controlled trials involving the police to develop a grounded theory on the factors contributing to high levels of treatment integrity in Police Field Experiments.

Peter William Neyroud

Wolfson College, University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee

Signed:

Peter William Neyroud

Dated: June 15th 2017

Abstract

Evidence-based policing (EBP) has emerged as a key strand of police innovation since Sherman's (1998) Police Foundation lecture. However, for others EBP raises as many questions as answers. One of the most contentious areas is the role advocated for randomised controlled trials in testing practice and developing knowledge to support EBP. RCTs are controversial with some scholars who argue that policing is not comparable to medicine and that RCTs are unable to reflect the complexity of the police role and context. Even those who advocate the use of RCTs recognise that there are significant challenges in achieving the high dosage and high fidelity that a successful experiment requires.

This dissertation responds to these challenges by analysing the completed randomised controlled trials in policing and using a case study, Operation Turning Point, to identify the factors that may contribute to the conduct and management of police field trials with high levels of treatment integrity. In the introduction, Chapter 1, the approach is set out, framed around grounded theory, to be developed in four, linked, chapters.

Chapter 2 is focused on understanding treatment integrity in RCTs involving the police: A search for police RCTs is produced 122 Police RCTs completed and reported by 2016. The levels of treatment integrity are analysed. 78 of the 122 RCTs exceeded a 60% threshold, with 49 being above 90%.

In Chapter 3, a "novice theory" is developed and tested as an explanation for levels of treatment integrity in police randomised controlled trials: Analysis of the 122 RCTs suggests that "novice theory" can provide an explanation for the general patterns of treatment integrity. Further detailed analysis suggested that there are, however, other factors which may be important in determining the treatment integrity.

These are developed in Chapter 4, which centres on a case study of Operation Turning Point. Using published case studies and an analysis of

juvenile justice RCTs, a potential framework of operational factors is developed that appear to be important in effective conduct and management. The Turning Point case study is used to develop and expand on those operational factors.

Finally, taking the two together, the analysis concluded that, beyond the operational factors, there were some more strategic, "protective factors" that were also critical. These are developed in Chapter 5, by using the coding and analysis of interviews with a sample of key staff involved in Turning Point

Our analysis suggests that novice theory needs to be understood in the context of both the operational and protective factors that we have identified. Taken together these findings indicate the potential advantages of building institutional frameworks in which the development of practitioners and researchers and the conduct and management of experimental research could be brought closer together. We conclude with ten recommendations designed to improve the treatment integrity of police RCTs.

.

Acknowledgements

I want, above all, to thank my supervisor, Professor Lawrence Sherman, who has been both an inspiration and a goad in my journey into experimental criminology.

The research was born in the custody suites and meeting rooms of West Midlands Police, where my role as research manager in Operation Turning Point sparked my curiosity about randomised trials in policing. The Chief Constable, Chris Sims and the staff in Operation Turning Point showed how police science can help to transform policing.

Mark Woodruff from the Monument Trust was best research funder you could have wished for – both demanding and passionate about the outcomes.

Along the way, I have been very fortunate to have the support and guidance of many distinguished scholars – notably Professors David Weisburd, Daniel Nagin, Sir Anthony Bottoms, David Farrington, Anthony Braga, David Wilson and Lorraine Mazerolle. I was also very grateful for the advice of my first-year reviewers – Professor Jonathon Shepherd and Dr Justice Tankebe.

I would also like to thank all the scholars who responded to my questions about their research and their randomised controlled trials.

My colleagues on the Campbell Crime and Justice Group and Police Executive Programme at Cambridge have been a constant source of support.

This Thesis is dedicated to my wife, Sarah and our four children – Eleanor (also a PhD candidate in Criminology), Huw, Florence and George – who have put up with me over the six years and more that this piece of research has taken to complete.

Table of Contents

List of Tables, Figures and Appendices

Figures: **Page**

Appendices:

1. Introduction:

1.1 Evidence-based policing and the randomised controlled trial in policing:

Public Policing in the developed world has been facing a “perfect storm” of challenges from a combination of fiscal austerity, the changing nature of crime and policing demands, new technology and loss of legitimacy from high profile critical incidents (Neyroud and Weisburd, 2014, Neyroud, 2014 and President’s Task Force on Policing for the 21st Century, 2015). Following Sherman’s Police Foundation lecture on “Evidence-based Policing” (1998), a movement to reform policing from within through the deployment of scientific approaches – “evidence-based policing” (EBP) – has grown in importance, at least partly in response to these challenges.

Sherman’s conception of EBP proposed an approach to policing for which he drew heavily on the experience of evidence-based medicine. Even though there were clearly differences between medicine and policing, Sherman (1984) had earlier argued that one key similarity between policing and clinical medicine that could be observed was the need for both to do “something about a problem, even if the something is merely likely, but not certain, to be helpful” (p.74). He suggested that police practices could be divided into three types: those based on ignorance (no evidence), on equity (the evidence supported neither one approach nor another) or on differentiation (where there was a clear preferred approach supported by the evidence), depending on the extent of the knowledge base supporting action.

Building on this typology in 1998, Sherman defined EBP as “the use of the best available research on the outcomes of police work to implement

guidelines and evaluate agencies, units, and officers (Sherman, 1998: 3). He contrasted evidence-based approaches with knowledge based on unsystematic experience and argued that experience should, instead, be used as the basis for hypotheses that could and should be tested in the field by methods including (but not exclusively) randomised controlled trials (RCTs).

Sherman further developed these ideas in his review of the "rise of Evidenced-Based Policing" (Sherman, 2013). He documented how, since the 1998 lecture, EBP had developed: how both the research base and the institutions to support the approach had continued to grow. He suggested that EBP could now be framed against three policing tasks, the "Triple T" of targeting, testing and tracking. Drawing on the Maryland scale of scientific methods (Sherman et al., 1997), he encouraged police to conduct their own tests of their practices using the highest standard of tests possible, with a strong preference for RCTs wherever the method was possible.

Sherman's call for RCTs has been mirrored in the direction taken by public policy in the UK. Haynes et al. (2012) boldly asserted "randomised controlled trials (RCTs) are the best way of determining whether a policy is working". They went on to lament that RCTs were not "routinely used to test the effectiveness of public policy interventions in the UK. We think that they should be" (p.4). Smith (2012) still complained that the Home Office had, as result of Chitty's (2005) review of evidence on "what works?", established a clear preference for RCTs. The Home Office review and the BIT paper have clear parallels in developments in US policy-making at Federal level where the Coalition for Evidence-based Policy (2015) has argued that funding should be tied to "top-tier evidence" supported by RCTs. Equally, at State level, the Washington State Institute of Public Policy has developed a framework for assessing programmes for their impact and cost-effectiveness, which places a strong reliance on RCT derived evidence (Washington State Institute for Public Policy, 2016).

This dissertation is focused on RCTs in one specific area of evidence-based public policy: policing. Taking up the challenge from Sherman (1984, 1998

and 2013) and Haynes et al. (2012), the starting point is an assessment of how many RCTs have been done in policing since the 1950's. Despite the growing interest in experiments, there is no systematic register of police experiments. There have been several searches for them either to provide resources for practitioners (Lum et al., 2011) or to explore the network of scholars and the social capital necessary to sustain experimentation (Braga et al., 2014). This research has drawn on the earlier searches and supplemented their efforts by drawing on the network of scholars, the journals that have published similar studies and the Grey literature. The aim has been to discover as wide a range of studies as possible to support an analysis of the factors that are associated with experiments which successfully achieve high levels of treatment integrity.

Sherman (2010) has been clear that it is important to separate a "successful" experiment from a "successful outcome". The researchers' success should be measured by the extent to which they conduct and manage an experiment that provides an internally valid test of the hypothesis under scrutiny (Torgerson and Torgerson, 2008), not the extent to which the outcomes of the tested intervention demonstrate high levels of effect size, cost-effectiveness or even statistical significance. This separation of the "success" of RCTs into two distinct parts - the conduct of a successful scientific test and the success of the intervention as judged through the measurement process of the test - is central to this research, which is focused on the former, rather than the latter.

Given this important distinction, the first half and most quantitative part of this dissertation – in Chapters 2 and 3 - is concerned with the extent to which the completed police experiments have achieved high levels of treatment integrity, arguing that this is a key measure of their successful conduct. Treatment integrity is defined as an estimation of the extent to which cases that have been assigned to a condition remain so, combined with the extent to which the key elements of the tested treatment were delivered. The second part of this quantitative section – Chapter 3 - develops and tests a "novice theory" from the analysis of treatment integrity.

Drawing on this data and a case study, Operation Turning Point, the second half and more qualitative part of the dissertation explores firstly the challenges of implementing high levels of treatment integrity in RCTs in policing and, then, with data from participants in Turning Point and from the completed trials, the ways in which successful experiments can be delivered effectively. The findings and implications of these two, linked halves are progressively drawn together to outline a grounded theory (Glaser and Strauss, 1967) as to how to conduct and manage successful RCTs in policing.

Throughout this dissertation there are two main sources of data: first, the reports and articles setting out the process and findings of the 122 completed randomised trials in policing; second, Operation Turning Point, a randomised controlled trial which tested the relative effectiveness of prosecution against a deferred prosecution with conditions in Birmingham, UK, between 2011 and 2014. The author was the principal investigator and research manager of this trial

The data on the completed trials is presented in Appendix 1. This presents key details of all the discovered RCTs which had been completed by 1st January 2016, the cut-off date for the search. Alongside the author, date and summary description of the trial, Appendix 1 provides the summary data on the sample size and estimated treatment integrity which has been derived, as set out in Chapter 2 below, by constructing a CONSORT diagram for each experiment. In a small number of these experiments there were existing CONSORT diagrams. However, given that the discipline of publishing a CONSORT is a relatively recent requirement in social sciences, most of the CONSORTs have been drawn up in the original form by the author for this dissertation from such details as are available in the published material.

The data from Turning Point includes a range of primary and secondary materials, but in particular: the CRIMPORT Protocol (Appendix 4); more than seven hundred documents from meeting minutes, emails and notes; interviews with a sample of 18 officers and staff who participated in the

research. There have also been twelve articles and chapters published to date reporting various aspects of the Turning Point research:

- Sherman and Neyroud (2012): the overall theory supporting "offender desistance policing" and the approach to testing it in the field.
- Neyroud and Slothower (2013): the interim report of Operation Turning Point.
- Neyroud (2014): the development of a triage system to discriminate between high and low harm offenders in police custody and the approach to testing it.
- Slothower, Sherman and Neyroud (2014): the findings from Operation Turning Point on the importance of a "training, tracking and feedback" model of implementation.
- Slothower (2014a): the findings from the victims' RCT in Operation Turning Point.
- Slothower (2014b): the findings on the importance of decision support systems to support consistent police discretion in out of court disposals.
- Bedford and Mazerolle (2014): organisational learning from Police RCTs.
- Neyroud and Slothower (2015): the challenges and opportunities in reforming police use of out of court disposals.
- Hobday (2015): targeting reasons for rejecting random assignment in the Turning Point RCT.
- Neyroud, E. C. (2015): the victim offender overlap in the Operation Turning Point sample of offenders.
- Neyroud (2017) on the ethics of practitioner led police research.
- Bedford and Neyroud (2017) on organisational learning from RCTs.

An analysis of the outcome data two years on from the completion of the experiment was also in progress at time of writing.

Drawing on this material, the main analysis and discussion is set out below in four linked chapters. However, before moving to these sections, it is important to frame the research by exploring four key issues (1) the reasons why RCTs in policing and public policy research are so important and, particularly, within the context of the movement towards evidence-based policing (2) the reasons why, despite this, Haynes et al. (2012) are right to observe that there have been relatively few of them in policing (3) why RCTs are controversial in policing (4) why paying attention to their effective conduct matters to the field of police research and to the development of evidence-based policing. Taken together these four issues provide the justification for this dissertation as a piece of research focused on RCTs in policing. They also help to frame the research questions for the four chapters, to which we will return below.

1.2 Evidence-based policing: the development of a movement and the importance of RCTs:

In the years since the 1998 lecture, an EBP movement has developed (Lum and Koper, 2017). This can be seen as part of a broader movement for evidence-based approaches within criminal justice and social policy (McGloin and Thomas, 2013), within which there are some distinctive elements in policing. Some key features that deserve to be highlighted are:

- A focus on "experimental criminology" as the spearhead of evidence based crime prevention and policing (Sherman, 1998, Sherman et al., 2002, Perry et al., 2006, Sherman, 2009 and Sherman, 2013). This has, for example, seen the development of the foundation and growth of the Academy of Experimental Criminology in 1998 and the establishment of the Journal of Experimental Criminology in 2005;
- Arguments supporting the potential of a broad discipline of "police science", with "ownership" from within the police as a mechanism to transform policing (Weisburd and Neyroud, 2011, Neyroud and Weisburd, 2014(a) and (b));

- The expanded use of systematic reviews of experimental and quasi-experimental studies in order to create a “force multiplier” of knowledge in policing (Weisburd and Telep, 2015). “Force multiplier” was derived from military doctrine and referred to a combination of assets or tactics that, in combination, increased the impact of the individual components. In this case, systematic reviews provided a means of enhancing the impact and external validity of a body of individual studies on the same intervention. The single most important development in this area has been the setting up of the Campbell Collaboration in 2000 and the Crime and Justice Group of Campbell, which has overseen the completion and publication of 25 policing relevant systematic reviews (Welsh et al., 2013, and Campbell Collaboration, 2016);
- The emergence of a new discipline of leadership and management centred around evidence as, for example, set out in Sherman’s Triple T approach (Sherman, 2013) and advocated by the Center for Evidence-based Management (Rousseau, 2012);
- The development of ideas for a reformed “profession” within policing. Manifestations of this can be seen in papers proposing a “new professionalism” in policing (Stone and Travis, 2011 and Neyroud and Sherman, 2013), the emergence of the UK College of Policing as a “professional body for policing” (Neyroud, 2011) and the creation and expansion of the Society of Evidence based Policing (SEBP) (in the UK, Australasia, Canada, the USA and Scandinavia);
- Connected with this has been pressure for a better qualified and “Chartered” Profession with evidence as a key underpinning factor (Neyroud, 2011, Neyroud and Sherman, 2013 and Canadian Council of Academies, 2014);
- The growth of professional police officers carrying out their own field research and research partnerships dedicated to police research (Sherman, 2013, Engel and Henderson, 2013 and Alpert et al., 2014)

Some have argued, despite these developments, that EBP, far from a "movement" or even a programme, should be seen, instead, as a tactical approach, narrowly focused on questions of "what works?" which can be put alongside a long series of other innovations such as intelligence-led policing, zero tolerance policing or problem-oriented policing (Greene, 2014, van Dijk et al., 2015 and Sparrow, 2016). However, the emerging institutional and professional changes set out above suggest that a more accurate assessment is Sherman's conception of EBP as a broader reform movement underpinned by the "belief that greater use of research could help transform policing into a more legitimate and respected profession" (Sherman, 2013:5).

Whether it is a tactical or a strategic reform, a central and controversial feature of EBP is the importance attached to the testing of key tactics and strategies and, especially, where appropriate, by randomised controlled trials. Sherman (1992, 1998, 2010 and 2013) and Weisburd (2004), Weisburd and Hinkle (2013), Lum et al. (2014) and Lum and Koper (2017) have argued that RCTs are essential to the evaluation of the effectiveness of key tactics and strategies in policing.

Weisburd and Hinkle (2013) put forward three main arguments for this stance. Firstly, the agencies engaged in crime prevention, such as the police, have to operate in an environment of constrained resources. As such they should seek to rely on the "most rigorous methodology" possible to inform their decisions. Secondly, with the complexity of the field there is significant risk in relying on evaluation methods that can be confounded by other factors, which may bias the research findings. Thirdly and in contrast, in an RCT, the process of random allocation reduces the likelihood that omitted variables are systematically related to the treatment and control sample. Researchers can, therefore, more safely assume that the only systematic difference between the treatment and control samples arise from the effects of the studied treatment (Shadish et al., 2002). Added to this, Lum et al. (2014) have argued that RCTs have tended to have a greater public policy impact than other types of research. Taken together, these authors have suggested that if the goal of

research is the advancement of evidenced based practice, the use of RCTs has much to commend it, both in terms of methodology and impact.

1.3 Why so few RCTs in Policing?

However, RCTs in the crime and justice field are still relatively rare, compared to fields such as medicine and education. The Cochrane Library already contained more than 150,000 randomised controlled trials in medicine by 2004 and more than half a million RCTs have been completed in that field (Goldacre, personal communication, 2016). In contrast, Mackenzie (2010) demonstrated that in the field of Corrections there were 284 quasi-experimental studies but only 42 RCTs. MacKenzie's figures may well have been an underestimate. A more recent search for the Global Police Database project (Higginson et al., 2015) has identified more than 7000 quantitative studies within policing since the 1950's. However, even with this much wider search, amongst those 7000 or more, there is a much, much smaller number of RCTs. When Braga et al. (2014) reported a systematic search for police RCTs, they could only find 63.

There are several reasons for this small number: ethics; feasibility; communication; academics as "critics" or "partners"; organisational culture. We will deal with these in turn.

1.2 (a) Ethics

First and foremost, there are undoubted ethical challenges, such as the provision of different treatments based solely on random assignment, which is often bound up with questions of about the legality of police withdrawing treatments or providing treatments only to selected subjects. The Federal Judicial Center (1981) identified a series of questions that need to be addressed to justify experimentation in justice: the practice must either need substantial improvement or be of doubtful effectiveness; there must be significant uncertainty about the value of the proposed innovation; there must

be no other practical means of resolving the uncertainty; the experiment must have a serious intent to inform future policy choices. Above all, as Goldman concluded “the greater the harms [from the departure from equal treatment] the greater the anticipated benefits in order to warrant an experiment” (1983: 734).

Dunford (1990) discussed the process of addressing these issues in persuading agency staff and local political leaders to conduct a RCT of diversion as part of a National Diversion trial (Dunford, Osgood and Weichelsbaum, 1982). The Federal Judicial Center questions provided the basis for arguing that the importance of questions being explored outweighed the concerns about equal treatment. In a second example, Dunford showed how the questions were answered after those involved in the trial agreed that using random assignment could be used to ration the limited resources available for a new treatment. In a final example, a domestic violence experiment that was one of a set replicating the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984), the experiment was designed around police officers field discretion to determine appropriate resolutions. Dunford (1990) concluded by advising researchers to work through the ethical and legal implications of an experiment long before they present their proposal to a potentially reluctant agency leadership. However, his analysis makes clear that the ethical challenges are not to be lightly dismissed. Goldman (1983) goes further and suggests that experimentation can only be justified over other methods as result of the greater likelihood of probative knowledge.

1.2(b) Feasibility

Farrington and Joliffe (2002) identified a second major obstacle: RCTs are not always feasible, even if the ethical challenges can be overcome. Their paper for the Home Office identified that even where the design would be highly desirable to determine the relative effectiveness of a treatment, context and case flow may render that ambition unachievable. Strang (2012) also

highlighted the problems of getting cases in a 'trickle-flow' RCT, even after the decision to embark on the trial. Moreover, the solution is not, in Perry et al.'s (2010) analysis, a simple matter of expanding the scope or timescale of the experiment, because this runs further risks to internal validity by compromising control of the study. On the other hand, Feder et al. (2000), Sherman (2010) and Strang (2012) pointed to the importance of creating a "coalition" to overcome hurdles with case flow and control of the experimental conditions and the need for a willing partner in the police to make this happen.

However, finding a willing police partner has not always been straightforward. There has, historically, been a general reluctance to support field research from within policing. Commenting on this, Weisburd and Neyroud (2011) argued "Most police agencies do not see science as critical to their everyday operations. Science is not an essential part of this police world. At best it is a luxury that can be useful but can also be done without." (p.3). They also suggested, from personal observations on either side of the police-researcher divide, that Chiefs, who operated within a highly political environment, tended to see more risks than benefits from testing innovations with designs like RCTs. The extent to which a RCT will be feasible is, therefore, determined by both a topic and context capable of experimentation and an organisational willingness to participate – an issue to which we will return in Chapter 5.

1.2(c) Communication

Yet, reflecting on the obstacles to effective police-research partnerships, Bradley and Nixon (2009) likened the relationship between researchers and police officers as a "dialogue of the deaf". In their analysis, the police and the academic community have fundamental disagreements about the purpose, process and outcomes of research. On the one hand, a large part of the existing body of research on policing has been from a critical tradition, which, as Cockbain and Knutsson (2014) argue, has tended to view policing as both marginal to crime reduction and potentially racist and oppressive (Galliher, 1999). This approach stands in stark contrast to what Police Chiefs are

interested in gaining from research. An early pioneer, Wilson (1950), advocated, that police should invest in research that supports practice, policy and strategy development.

1.2. (d) Academics: critics or partners?

At the “critical” end of this debate, Hirschi (1993) condemned the proponents of research on the effectiveness of the police as being mere “administrative criminologists” (p.349), who practice a “practical criminology” (p.350), which is short on theory and uncritical of the role of government in addressing crime problems. Instead, Hirschi argued that research must be informed by human rights values and present a moral vision independent from government and institutional funders. Hirschi (1993) described crime as a “moral problem that is beyond the reach of experimentation” (p.350). In effect, Hirschi suggests that research in criminal justice should be the preserve of scholars who are independent from police or government rather than a partnership between researchers and the police focused on informing and developing police practices.

It is unsurprising, given the polar opposite positions, that the two main traditions of police research – “critical” and “police policy research” (Bradley and Nixon, 2009) or “applied police research (Cockbain and Knutsson, 2014) – have very different approaches to research and its outcomes (Manning, 2005). The former seeks to examine the organisation and its impacts as a means of holding the police to account, the latter to “develop theories, frameworks and/or empirical evidence to inform and support policing policy and practice” (Cockbain and Knutsson, 2014:2). As a result of the very different starting points, the methods, style and philosophy of research have tended to be very different. Critical researchers have prided themselves on detachment and independence from the police. The focus of the research is to provide an expert voice to inform the thinking of the citizen and those charged with the governance of the police (Bradley and Nixon, 2009).

Police policy research, on the other hand, is committed to much closer engagement and partnership with the organisation, with a view to improving and developing policing policies and practices based on evidence. Its advocates have argued that the "threat of less objectivity is outweighed by the gains in better understanding of contextual influences" (Engel and Whalen, 2010: 111). Moreover, as Greene (2014) demonstrated, applied police research in no way implies that the research is uncritical or unable to explore and report uncomfortable findings.

The design and delivery of RCTs in policing places them firmly in the police policy rather than critical tradition of police research. On the one hand, the central focus and rationale for conducting an RCT is "addressing the 'what works?' question in 'evidence-informed' policy-making and practice" (Torgerson and Torgerson, 2008:1). On the other hand, the requirements for implementing the design described by Sherman (2010) requires a trusted relationship between the researcher, the agency involved and the practitioner community, which is in clear contrast to Hirschi's independent scholar model.

The range of different approaches to researching the police goes some way to explaining the small number of RCTs. Experimental criminology and the conduct of field experiments in policing has been and remains a specialism within an already narrow field, policing research within criminology. As Braga et al. (2014) demonstrated a very small group of scholars were responsible for the vast majority of the RCTs that they found. Only 11 authors had more than one study to their name and only four, Sherman, Strang, Davis and Weisburd, had been the lead authors for 5 or more. In marked contrast to what one would expect to find in medicine, there were only three police practitioners listed as authors or investigators and no other police staff feature in the longer author list of 126 (see Braga et al., 2014: Table 3).

1.2. (e) Police Organisational Culture

Shepherd (2003) sought to explain the “famine” of police RCTs by comparing medicine and policing. He contrasted the organisational culture and status of operational research in the two professions. He highlighted the fact that medical interventions have been largely the responsibility of “clinical practitioners”, whereas police RCTs evaluating policing interventions have been almost exclusively conducted by academics who have no clinical, policing experience, let alone current practice. Indeed, in support of Shepherd, the National Health Service website in the UK states “Research and clinical trials are an everyday part of the work done in the NHS. The people who carry out research are mostly the same doctors and healthcare professionals who treat people” (NHS, 2015). Shepherd also suggested that the fact that medical education and medical research are both “advanced in, or closely associated with, university teaching hospitals” contrasted strongly with the “absence of university police schools” (p. 290). Weisburd (2003) described this situation as a failure to develop a “comprehensive infrastructure for experimental evaluation” (p.336). As a result, RCTs in policing, based on Shepherd (2003) and Braga et al.’s (2014) analysis, have been a specialist academic pursuit, with a small range of funders and reliant on a series of largely one-off partnerships being built between university-based researchers external to the police and a police agency.

Alpert et al. (2013) evaluated the very considerable difficulties faced in the building and sustaining such partnerships. Rojek et al. (2015), reflecting on the earlier evaluation, were optimistic that the field was changing. They suggested that police had been the “unwitting recipients” of research in the past but that there had been an important, recent shift in police officer education. The greater number of officers with tertiary qualifications “has led to incremental change for the better” (p. 74). Rojek et al. suggested, much as Shepherd had done in 2003, that this offered an opportunity for police officers and their forces to “engage in properly designed and supervised longitudinal research and randomised controlled trials” (p.74). Nevertheless, Braga et al. (2014)’s analysis of police RCTs prior to 2012 suggested that Sherman’s (1979) earlier advocacy of “research police agencies” or subsequent (Sherman, 2010) proposal for “field stations” had remained largely unfulfilled.

1.3 Difficult and controversial? RCTs in Policing:

As the field changes and the opportunity for more RCTs and potentially "clinically-led" practitioner RCTs opens up, it is all the more important to recognise that even after 50 years of conducting them, RCTs in general (Deaton and Cartwright, 2016) and RCTs in policing specifically remain both difficult to do and controversial (Sparrow, 2011 and 2016). The debate has generally been focused around four connected issues (1) whether RCTs can be seen as a "gold standard" or "best method" of evaluating effectiveness when compared to other research methods (2) the ability of police field experiments to meet acceptable levels of treatment integrity – the internal validity of the experiment (3) the extent to which the outcomes of individual experiments in one area can be generalised to other contexts – the external validity (4) a fierce debate as to whether the RCT design does, indeed, allow researchers to draw strong inferences as to causality by comparing two matched samples separated only by the random assignment of treatment provision.

1.3 (a) The Gold Standard?

As to whether RCTs are the "gold standard", Weisburd (2003) argued that there should be an ethical preference for RCTs because they provide "valid answers to questions about the effectiveness of treatments" by "ruling out alternative causes of the outcomes observed" (p.350). Citing these same reasons Torgerson and Torgerson (2008) unequivocally stated that RCTs are the "'gold standard' research method for addressing effectiveness questions in health, education and social policy" (p.8). As Farrington (1983) noted, Weisburd and Torgerson and Torgerson's argument relied on experiments being implemented with the full dosage and high fidelity to meet the requirements for internal validity. RCTs "unique advantage over other methods is their high internal validity, or high ability to demonstrate the effect of one factor on another" (p.257).

1.3. (b) Internal validity

Berk (2005), whilst identifying himself as a supporter of the method, criticised the "rhetoric" and the overselling of RCTs and relegated them to the "bronze standard". His main reason for this was that the "most serious vulnerability of randomised experiments is their implementation. It is difficult to do experiments well" (p.429). He cited five potential problems, all related to the internal validity of the experiment: achieving a sufficient and timely case flow; the level of attrition from the sample; the failure to implement random assignment; the mis-assignment or re-assignment of treatments within the experiment; poor measurement of key variables. Sherman (1992) and Feder et al. (2000) supported this analysis. They argued that police and criminal justice field experiments require a particularly disciplined approach to achieve successful implementation and high levels of internal validity. So much so that Sherman (2010) identified a long and demanding list of prerequisites, drawn from experience in conducting experiments in domestic violence, hotspots policing, repeat offender strategies and restorative justice.

1.3. (c) External Validity

Even when successfully completed, some have argued (Cartwright and Hardie, 2012, Sampson et al., 2013 and Laycock and Mallender, 2015) that there are "many RCTs which admirably demonstrate internal validity but fail abysmally when considering the external validity" (Laycock and Mallender, 2015:7). They point back to the medical analogy, suggesting that "crime, and criminal justice are social phenomena" (Sampson et al., 2013:20) which are highly contextual and less easily generalizable than a medical intervention. Eck (2002) identified what he called the "diabolical dilemma" (p.104): the more that police field experiments are tightly controlled, the less that they represent a "real world" of loosely controlled operations. Equally, it seems that the greater the emphasis on replicating real world conditions, the greater the

likelihood of compromises to the tight requirements necessary to achieve internal validity (Maxfield and Babbie, 2015).

1.3. (d) Causal Inference

In order to overcome Eck's "diabolical dilemma", Shadish et al. (2002) proposed that a firmer emphasis should be placed on internal validity when seeking to establish cause and effect. However, even this pragmatic balancing acts runs up against the fourth area of debate. For, causality is at the heart of one of the most important debates about RCTs. Pawson and Tilley (1997), christening themselves "realistic evaluators", criticised the "experimentalists" because they have "pursued too single-mindedly the question of whether a program works at the expense of knowing why it works" (p. xv). They have made a convincing case on the need to pay attention to mechanisms, but not on the capability of well conducted experiments to produce "knowledge" that is "likely to more probative" (Goldman, 1983).

These debates are particularly important in policing research. Maguire et al. (2015) found, in surveying police reforms and their evaluations for a study of problem-oriented policing, that "the result is often shallow or incomplete implementation (low dosage) or inaccurate or unfocused implementation (low fidelity)" (p.72). Similar findings can be found in other areas of research into the implementation of police innovations such as team policing (Sherman et al., 2014 and Walker, 1993), problem solving and hot spotting (Irving and Dixon, 2002 and Sherman et al., 2014), community policing (Vito et al., 2005), mental health diversion schemes (Reuland, 2004) and domestic violence experiments (Sherman, 1992). As a result, of the four areas of debate presented here, internal validity stands out as of particular concern. Yet, Eck's "diabolical dilemma" is also highly relevant. The researcher's attempts to overcome internal validity problems by tightly controlling the environment run the twin risks of reducing generalisability and diminishing support from the operational field, long used to the dictum that "police experiments never fail". This is not a methodological truth, but rather a weary and somewhat cynical

observation that has greeted many top-down innovations in an organisation whose staff perceive a tendency to shift from one innovation to another (Weatheritt, 1986 and Weisburd and Neyroud, 2011).

1.4 Effective conduct and management of Police RCTs:

Policing is a challenging environment in which to conduct RCTs. Yet, as we have seen, RCTs are critical to the development of evidence-based policing. It is, therefore, surprising that, until recently, there has been relatively little systematic focus on Police RCTs as a collection of studies. The key problem has been that there is no register of police RCTs. Every scholar wishing to undertake such a collective analysis has first had to search for the studies. There have been a number of such searches over the last 25 years for criminal justice or crime reduction RCTs (Dennis, 1988, Weisburd et al., 1990, Petrosino, 1998 and Farrington and Welch, 2006). These searches have included a number of police RCTs (see Table 2.1). More recently, Lum et al. (2011) have developed a comprehensive register of crime reduction related Police RCTs and Quasi-Experiments for the George Mason Matrix of Evidence-based Crime Reduction. Drawing on this Matrix and a wider search, Braga et al. (2014), in the most systematic search thus far, found and analysed 63 Police RCTs in order to explore the research networks and production process.

Braga et al.'s (2014) study has demonstrated the potential and importance of such an approach to developing our understanding of the field. However, the focus of their analysis was on the research process and the academic networks rather than on the broader issues of the conduct and management of RCTs in policing. For this, we have to look to a small number of studies by the group of Police researchers identified by Braga et al. (2014). The most important of these have been Sherman's detailed analysis of the Domestic Violence RCTs from the 1990's (Sherman, 1992) and Sherman and Strang's lessons from the Restorative Justice RCT's (Sherman et al., 2015). Sherman elaborated these in a key Chapter in the Handbook of Quantitative

Criminology (Sherman, 2010), which serves as a 'roadmap' for any researcher undertaking an experiment in criminal justice generally and policing in particular. To this can be added Strang's (2012) description of the frustrations of maintaining research collaborations through to the conclusion of a randomised trial.

Sherman and Strang's work is based on a detailed and grounded understanding of 25 years of experiments in a wide range of policing interventions. However, they were not able to draw on the whole corpus of police RCTs, because such a body of work has, up to now, not been readily accessible. Furthermore, Braga et al. (2014) excluded some studies before 2012 and all studies completed after 2012. Their cut off point and exclusions reduced the heterogeneity of subjects included. More importantly, it also meant that the development of practitioner –led or partnered RCTs, which have grown significantly since 2012, were not included in their analysis. The most obvious group of studies that have eluded their search have been the RCTs of Body Worn Video that, Lum et al. (2015) have shown, now number more than a dozen studies either wholly or partially complete. Many of these studies have involved practitioner researchers conducting field research for the first time (Ariel and Farrar, 2014 and Ariel et al., 2016). Many have also involved new experimental researchers who do not feature in Braga et al.'s tight network of 60 or so scholars (for example, Bradford and MacQueen, 2014). The widening of the group of experimental scholars and the inclusion of "pracademics" (practitioner-academics), professional institutes (such as the UK College of Policing) and Police Universities (such as the Norwegian Police University College) suggests that detailed analysis of the lessons from conducting and managing RCTs is ever more important.

It is also important for two other reasons, both of which are related to replication. Replication is not only a critical scientific principle, it is also a vital policy and practice challenge. It is one thing to carry out an experiment, quite another to repeat it in a different place and context and then generalise the findings and embed them into day-to-day practice. The Open Science Collaboration (2015) found it is not always rigorously pursued. The

Collaboration found when they replicated 100 psychological studies that in less than half the cases could they unambiguously find the same results. The Collaboration's approach was careful to involve the original teams in ensuring that the research designs and practices were consistent with their experiment. They were, therefore, able to reduce the impact of different approaches to conduct and management. In contrast, researchers engaged in police field experiments find themselves confronted by both the challenges of negotiating the experiment into the field (Dunford, 1990) and the complexities and heterogeneity of the field itself (Sampson, 2010). This suggests how important it is that police researchers can control the process of conducting and managing the experiment by drawing on practice and guidelines distilled from the experience of RCTs that have already been completed.

Such consistency is also important when considering the reliability of studies for systematic review (Higgins and Green, 2008). The Cochrane Collaboration requires every systematic review to estimate the risk of bias from the studies included in the review. The Cochrane assessment of the risk of bias focuses on the "risk of bias in the actual design and conduct of the study" rather than results reporting (Higgins and Altman, 2008: 8.3.2). There are some key differences between the medical RCTs that feature in Cochrane Reviews and the social science research that the Campbell Collaboration reviews. Above all, it is relatively rare, usually impractical and often unethical to "blind" a subject to an intervention in a criminal justice setting, whereas this is a standard requirement in most medical trials. However, once allowing for the differences, two particular features of study "conduct" are clearly relevant to both medical and police and criminal justice studies: attrition bias arising from problems with the process of randomisation; performance bias as a result of problems delivering the intended dosage.

Attrition and performance are often closely linked and controlling them through the effective conduct and management of the experiment is essential to maintaining high levels of "treatment integrity" in any RCT. Through this study "treatment integrity" as a combined estimate of the impact of potential

attrition and performance bias will be used as a key metric for the successful conduct and management of RCTs.

The central argument of this thesis is that it is possible to improve experimentation in policing through better understanding the lessons from the whole corpus of completed police experiments. In order to do this, we first need to find as many of the relevant studies as possible. Secondly, we need to understand the consistency and issues related to treatment integrity, as far as is possible from the reports and articles available. Thirdly, we need to see how far we can identify common issues in the way that Police RCTs have been conducted that may be associated with high or low levels of treatment integrity. Finally, we need a working, grounded theory about how to conduct successful, high integrity police experiments.

1.5 Overview of the Method, Structure and Research Questions:

1.5.1: Method:

The overall approach for this thesis was based around grounded theory, which, as Urquhart has set out, should be "designed to generate or discover a theory", which should consist of a "plausible relationship between concepts or sets of concepts" and can be reported "in a narrative framework or a set of propositions" (Urquhart, 2013:5). Glaser and Strauss (1967) described the approach as a practical rather than grand theoretical approach to developing theories from both qualitative and quantitative data.

Grounded theory should not, they suggested, ignore the existing literature but should, as far as possible, start from a "non-committal" stance. For, Glaser and Strauss (1967) advise, the researcher cannot approach reality as a "tabula rasa" (p.3). However, they subsequently adopted divergent approaches. Glaser (1992) encouraged distance from the literature to allow findings to emerge from the research. Strauss (1987) accepted that this was often unachievable. Urquhart (2013) suggests that it is often helpful to

approach it in phases, which become progressively more engaged in the literature, moving from open minded exploration towards integration.

It is Urquhart's phased approach that has been adopted. Such an approach not only fitted better with the natural phasing in the research programme but also with the much longer research time frame dictated by a part time PhD. Instead of the more normal cycle of three years, usually divided into three sequential phases of literature review, field work and thesis writing, this research has evolved over nearly seven years from 2010-2017. This has both allowed Operation Turning Point to be completed (2011-2014) and for the initial focus of the research to broaden from a detailed focus on the evaluation of one RCT to the conduct and management of police RCTs more generally. As such, the 'research journey' itself has been a gradual, grounded appreciation of a significant and increasingly important gap in our knowledge about police RCTs.

In practical terms this has meant that there have been four connected and sometimes parallel phases: an initial engagement with the literature on experimentation and police RCTs, which was both necessary to frame the initial research proposal and to support the conduct and management of the Turning Point experiment; the search for and preliminary analysis of completed police RCTs; the development and deployment of an interview protocol in order to interview a sample of staff involved in Turning Point; detailed analysis of the treatment integrity in police RCTs and of the case study of Turning Point.

The "novice theory", which has emerged and which is set out in detail in Chapter 3, was gradually developed using Urquhart's approach of a phased exploration starting with open coding and gradually developing a more detailed and, ultimately, more integrated theory: starting with a high level, open coding of the Police RCTs for factors that appeared to associated with treatment integrity; then a more detailed, focused coding of the 122 Police RCTs for the novice categories; utilising the open coded categories developed from the 122 Police RCTs to frame the interviews of a sample of participants

in the experiment; finally, a selective coding of the interviews, using HyperResearch for Mac (Hesse-Biber et al., 1991 and Lewins and Silver, 2007), which is presented in Chapter 5. .

Each chapter presents a more detailed methods section which describes the approach adopted for that particular chapter: Chapter 2 is based on a search for and descriptive analysis of the 122 Police RCTs; Chapter 3 develops "novice theory" from the focused coding of the novice categories from the 122 Police RCTs; Chapter 4 uses a Case Study approach to explore the operational factors that appear to contribute to high levels of treatment integrity'; Chapter 5 draws on the coded interviews of participants in the Case Study experiment to develop a model of "protective factors" that appear to support high levels of treatment integrity.

1.5.2: Structure and Research Questions:

The dissertation has been structured into four chapters, which explore these linked themes around the conduct and management of randomised controlled trials in policing. The starting point is provided by a search for the completed randomised controlled trials in policing. A full list of the trials found is presented at Appendix 1.

Starting with the data from the search, Chapter 2 presents an analysis of the trials focused on the level of treatment integrity apparent in the reported material. The research questions for this chapter are:

(a) What is the universe of RCTs in policing that have been completed and reported by 2016?

(b) What is the level of treatment integrity in completed and reported RCTs in policing?

The findings from Chapter 2 suggest that treatment integrity levels have been rising. However, it appears that there has been some recent slippage, which may be associated with an expansion of police RCTs involving new investigators, new departments and previously un-researched interventions. In Chapter 3 we explore the extent to which a "novice theory" can be identified as a key factor linked to levels of treatment integrity. The research questions explored are:

(a) To what extent can "novice theory" explain low and high levels of treatment integrity in RCTs in policing?

(b) What are the implications of these findings for the conduct and management of police RCTs?

Chapter 4 turns the question about treatment integrity around and uses a qualitative approach to develop the analysis. The Chapter is centred on a case study – Operation Turning Point – which is framed by analysing juvenile justice and case studies of police RCTs. Operation Turning Point was a randomised controlled trial in Birmingham, UK, comparing prosecution (control) with a deferred prosecution with conditions (treatment), which the author managed from 2010-2014. The Turning Point case study draws on the notes, meetings and observations from Turning Point to describe the steps taken to ensure high levels of treatment integrity in a trial that ultimately delivered high levels (over 90%) after some significant initial problems. The trial was conducted in four phases and it was only in the final, fourth stage that high levels of treatment integrity were achieved. The case study is focused on the following research questions:

(a) What lessons for achieving "high integrity" can be derived from an analysis of published case studies of police RCTs and analysis of completed juvenile justice RCTs?

(b) Using the framework from that analysis, what can a case study analysis of Operation Turning Point, a "high treatment

integrity" experiment, tell us about the factors that might contribute to the conduct and management of successful (above 90% treatment integrity) randomised controlled trials in policing?

(c) **How might those factors add to or modify the "novice theory" proposed and analysed in Chapter 3?**

The final Chapter, Chapter 5, continues the analysis of Operation Turning Point but turns to the experience of the police staff involved in a randomised controlled trial. One finding from the analysis of the 122 RCTs is that there has been a significant shift to practitioner led or, at the very least, practitioner managed RCT research. In Turning Point, the author, as a lead researcher, was a practitioner and there were several practitioners involved in research side of the RCT. West Midlands Police, the host agency for Turning Point, was also an example of Sherman's "field station" - a force that had committed to a programme of RCTs. Finally, Weisburd and Neyroud (2011) advocated the importance of "ownership" and the development of a professional discipline of police science, the role, contribution and experience of police staff engaged in a randomised controlled trial.

The Chapter draws on the findings from interviews with staff involved in Operation Turning Point to explore these and other "protective" factors that may contribute to high treatment integrity. The Chapter uses a grounded theory method of open coding and theoretical sampling to construct a descriptive analysis of these themes and then compare them to analysis from the Matrix of experiments in Chapter 2, 3 and 4. The research questions explored are:

(a) To what extent is it possible to construct a model of the key protective factors that contribute to high levels of Treatment integrity in RCT field experiments in policing?

(b) **How might those factors add to or modify the "novice theory" proposed and analysed in Chapter 3?**

The final section, the "conclusions and recommendations" draws the three chapters together and sets out recommendations for future research and development of the field of experimental criminology in policing.

2. Understanding and Analysing Treatment Integrity in randomised controlled trials in policing:

2.1 Introduction: "The Controlled Trial in Institutional Research":

Clarke and Cornish's (1972) critique of "controlled trials" in "Institutional Research" has proved one of the most important and lasting contributions to the debates on the use of Randomised Controlled Trials (RCTs) in criminal justice settings. Their monograph was published at a time in the early 1970's when RCTs had been gaining ground as a research method for evaluation research in social science (Campbell and Stanley, 1963). Rose and Hamilton (1970) had just published their RCT in the British Journal of Criminology comparing different treatments in police cautioning practices in a British police force. However, as a result of Clarke and Cornish's paper (Farrington, 2003), it was to be 30 years before the next UK based RCT in policing was completed - Strang and Sherman's Restorative Justice experiments (Shapland et al. 2006). Petrosino et al. (2006) found that the British Journal of Criminology did not publish any UK RCTs in the intervening years and only one US-based RCT study. This seems to have reflected both the 'chilling' impact of Clarke and Cornish in the UK and the continuing controversy around the RCT method.

Part of the impact of Clarke and Cornish's paper can be ascribed to a wider shift in the official Home Office attitude to RCTs (Smith, 2012). Part can also be attributed to Clarke's development of "crime science" as a new discipline, with its own group of disciples (Clarke, 2004) and the connected, subsequent development of realist evaluation (Pawson and Tilley, 2004). However, there was also real substance to Clarke and Cornish's critique, which, as is stated in the foreword by T.S. Lodge, the then Home Office Director of Research and Statistics, redressed "the unwarranted optimism in the literature about the potential of such studies" (p.iii). Their observations arose from their reflections, or better still, their frustrations, from the implementation of a RCT in the Kingswood Training School (Cornish and Clarke, 1975). They centred their critique on a number of significant implementation issues that they had

encountered, which were largely concerned with the problems of control and management of the random assignment and treatment process.

Clarke and Cornish broke the problems down to five issues: ethical problems and the consent of practitioners; the choice of success criteria; the practical difficulties of implementing random allocation; problems of generalising from results; problems arising from changes in treatment through the lifetime of the trial. At the heart of Clarke and Cornish's analysis was Eck's "diabolical dilemma" (Eck, 2002): securing support for the experiment and meeting the ethical challenges raised had, they argued, required the "real world" to be so bent to the RCT design that both wider generalization and the day-to-day operational management of the institution were compromised. Equally, the "real world" needs of managing the institution and the professional investment by some staff in the experimental treatment made maintaining high levels of treatment integrity difficult. Clarke and Cornish concluded that these difficulties, added to the problems of complexity and isolating the effective mechanisms of change in criminal justice setting, meant that the "controlled trial would seem to have a more limited function in penal research than has sometimes been ascribed to it in the past and certainly more limited than it has in medicine" (1972:21).

Whilst Clarke and Cornish saw the Kingswood RCT as grounds for choosing alternative research methods, it is equally valid to see their list of obstacles to effective conduct and management of RCTs as a checklist of challenges that any experimentalist must necessarily tackle in designing and managing RCTs in a criminal justice setting. Indeed, given that Clarke and Cornish were describing an experiment inside a closed penal environment, it is arguable that the issues that they raised are even more pertinent in the potentially less controlled environment of street policing or pre-court investigation and decision-making.

This would seem to be borne out by the travails of a number of police RCTs. Sherman's (1992) detailed analysis comparing the implementation of the Minneapolis and Milwaukee Domestic Violence experiments highlighted the

problems with treatment integrity experienced in the former and the extensive steps required to overcome them in the latter. Dunford et al. (1985) wrestled with all five issues with mixed success in implementing the National (U.S.) Diversion experiments. Whilst some of their sites were more successful than others, the researchers were forced to comment that "because some youths in the study did not receive the services to which they were assigned and because some youths received services from dispositions to which they were not assigned, this research design does not represent a pure test of diversion programs" (p.29). McCold and Wachtel's (1998) Bethlehem Family Group Conferencing Project also failed to overcome Clarke and Cornish's obstacles. A decision to randomise treatment assignment before consent was obtained produced such a low level of treatment integrity that it is difficult to interpret the results with any certainty.

It appears from these studies that two of Clarke and Cornish's five issues, both of which are critical to the internal validity of the experiment, may be more significant in police field experiments. The first of these is the implementation and control of the random assignment process. The second is the delivery and consistency of the treatment intervention. Maguire et al. (2015) have demonstrated that these two issues, which they called the problems of "low dosage" and "low fidelity", are more general problems with the implementation of interventions in policing. This finding chimes with the wider lessons of research into the implementation of police innovations such as team policing (Sherman et al., 2014 and Walker, 1993), problem solving and hot spotting (Irving and Dixon, 2002 and Sherman et al., 2014), community policing (Vito et al., 2005) and mental health diversion schemes (Reuland, 2004).

In the context of an experiment, this well documented difficulty in sustaining tight discipline in design and delivery translates into problems with the "treatment integrity" of the experiment. This encompasses the set of processes by which the sample is first randomly divided, then consistently sustained through the experiment and, finally, the treatment being tested is applied to the sample at a specified threshold. Sherman and Strang (2004)

have argued that it can be problematic to separate these processes in a field experiment and that they should, generally, be considered as one continuous process designed to ensure the "integrity of the randomised assignment procedure" (Boruch, 1997: 100) and, thereby, the internal validity of the experiment.

In the absence of high levels of treatment integrity, there is a well-recognized risk of "attrition" and "performance" bias, which be seen to cast doubt on the reliability of the results (Higgins and Altman, 2008). It is much more difficult in criminal justice settings to adopt procedures such as allocation concealment and blinding of participants to the treatment condition, which are routinely required in medical trials to reduce the risk of bias (Akobeng, 2005). This means that, even more in criminal justice and policing than other areas, researchers seeking to conduct "successful" experiments need to pay the most careful attention to high levels of treatment integrity. According to Boruch (1997) even a 10% to 15% rate of discrepancy between treatment as randomised and treatment as actually received should warrant further investigation.

Despite the number of studies identifying the critical importance of paying attention to treatment integrity in assessing the reliability of RCTs, there have been relatively few studies of the level of treatment integrity across all Police RCTs. Those that have been done have explored only the treatment integrity of a cluster of similar trials (Sherman and Strang, 2004 and Sherman et al. 2016). There has been no comprehensive analysis nor any register of police experiments upon which to draw for such an analysis. There has, however, been a steady and accelerating growth in the number of Police RCTs and the range of topics. Braga et al. (2014) were able to identify 63 in a search which included studies completed by 2012. Searching for this study, which is reported below, has identified a further 59 that have been completed before or since, with more than 60 "in-flight" (Appendix 1 & 2). This number still compares unfavourably with the fields of medicine or education. However, by 2014, the progress since Farrington and Welsh (2006) found 16 policing

studies and judged RCTs to be “relatively uncommon” (p.56) has been considerable.

This chapter seeks to fill a key gap in the recent literature by presenting an analysis of the treatment integrity in RCTs in policing completed and reported by the end of 2015. The chapter is set out in four sections: an overview of the existing literature relating to treatment integrity in RCTs and particularly in criminal justice and policing; the methods for the search and estimation of treatment integrity; presentation of the results of the treatment integrity analysis; discussion and implications from the results. The chapter is concerned with answering two research questions:

(a) What is the universe of RCTs in policing that have been completed and reported by 2016?

(b) What is the level of treatment integrity in completed and reported RCTs in policing?

2.2 The Problem of Treatment Integrity in Police RCTs:

Even from the “first generation of systematic field experimentation” methods, in the 1960’s and 1970’s, experts (Boruch, McSweeney and Sodertrom, 1978) have acknowledged the challenge of successful implementation and the consequent threats to internal validity from compromise to treatment integrity (Cook and Campbell, 1979: 385). Cook and Campbell provided “cautious advocacy” for RCTs at the end of that first phase which had seen a significant deployment of the design in criminal justice as well as in education, social welfare and law (Boruch, McSweeney and Sodertrom, 1978). Their cautions were centred on the challenges of successful implementation. On the one hand RCTs provided the potential for a high quality of causal inference, on the other they carried a threat of problems with internal validity that they argued were so serious that they could only be balanced by running a parallel quasi-experimental fallback in the event that the design failed. Drawing on the

"teething problems" of the early RCTs, they identified and proposed solutions to a range of problems that had emerged. Two serious threats dominated their analysis: attrition from randomised samples and compliance with treatment delivery.

Nevertheless, Cook and Campbell (1979) did not flinch from the views set out earlier by Campbell and Boruch (1976) that successful RCTs had such significant advantages over Quasi-Experiments in reducing potential measurement bias that their use was justified despite the challenges. Their advocacy came in the wake of a vigorous debate about "what works" in criminal justice research, which had been triggered, in the USA, by Martinson's (1974) review of the findings from experimental and quasi-experimental research in penal research. Martinson found that the data from these studies had not provided convincing evidence of effectiveness. Petrosino (1997) observed that most commentators have focused on Martinson's findings that "nothing works" rather than on his criticism of the quality of much of the research that he reviewed. For this study, Martinson's most important impact has been on the consequent search for improved methods and strategies to achieve more successful experiments. Cullen (2013), in reviewing that impact, traced a strong connection between the debate that followed and the development of the meta-analytical techniques supporting systematic reviews and the emergence of the "evidence-based" movement in corrections and criminal justice.

One of the key pre-requisites for that emerging movement was the need to build a systematic understanding of lessons from the experiments that had been completed. To this end, Dennis (1988) undertook a systematic search for criminal justice RCTs and then a detailed analysis of a sample of forty "field" experiments, nine of which involved the police, in order to understand the factors that contributed to their successful or unsuccessful implementation. He carefully selected "field experiments" as oppose to "laboratory experiments". Dennis started with the presumption, drawn from the growing literature on RCTs including Clarke and Cornish (1972), that field experiments are often "logistical failures" (Dennis, 1990: 347). From his

analysis he observed that experiments “can and often do fail” and that the “failures are often in their implementation and consequential validity of their statistical inferences” (Dennis, 1990: 348). In order to understand the reasons for failure he focused on the process of constructing the experiment and on assessing the treatment integrity of the forty RCTs. He then classified the issues that appeared to have compromised the integrity of the experiments.

Dennis was able to identify six threats, four of which were directly related to attrition or treatment delivery (1) treatment dilution (2) treatment contamination or confounding (3) inaccurate case flow and power estimates (4) violations of the random assignment process (5) changes in the environmental context (6) changes in the treatment regimes. For each of these threats he proposed and discussed potential solutions. However, like Cook and Campbell (1979), when confronted with significant violations of random assignment process he could only proffer the option of embedding the experiment inside a quasi-experiment as a means of mitigation. In effect, what his analysis and that of Cook and Campbell suggested is that incurring significant failures in treatment integrity means that researchers must accept that this is tantamount to a failure of the RCT design. In turn, they must then fall back on a potentially weaker, quasi-experimental design.

The implication of Dennis’ argument is that researchers should be much less concerned about the magnitude of the effect size in the outcome data in assessing the “success” of an RCT. Instead, the proper yardstick of the relative success or failure of the conduct of a RCT is the extent to which the implementation of the experiment ensured that the intended treatments are delivered in the intended dosage to those intended and assigned to them. Indeed, he specifically spelt out his conclusion out: “failure to find a statistical difference between two or more regimens is not a failure from a research point of view” (Dennis, 1990: 369) as long as that failure did not derive from either insufficient statistical power or “implementation problems”.

Berk (2005) came to similar conclusions in assessing the strengths and weaknesses of RCTs. Describing them as the “bronze” rather than the gold

standard, Berk identified RCTs' most serious vulnerability as "implementation" and, like Dennis, narrowed down a set of "especially nettlesome problems" (Berk, 2005: 429) including case flow (in trickle flow designs), attrition, mis-assignment or reassignment and failures of treatment delivery. For Berk, Smyth and Sherman (1988) the best way to tackle such implementation failures generally and treatment integrity implications in particular was to prevent them in the first place. They suggested five steps that needed to be taken: recruiting the right staff to participate; training them carefully; building and maintaining a relationship and mechanism to air and resolve problems; actively observe the key processes of the experiment; design the randomisation process in such a way as to minimize cheating. Furthermore, given the inevitability that some officers would still seek to circumvent the rules, it was also important to provide some overt means to allow exclusions to be made and monitored.

Berk, Smyth and Sherman (1988) proposed that researchers pay attention to the conduct and management of the RCT in order to prevent treatment integrity problems. They were drawing on the lessons from experiments such as the Minneapolis Domestic Violence experiment. However, given that Sherman and Berk (1984) had already experienced treatment integrity problems in the conduct of the experiment, the subsequent evaluation raised the question of how far it would be possible to overcome problems of treatment integrity that had occurred already by adjusting for them in the post-implementation analysis. Berk and Sherman found that they had to make "allowance" for "occasional alternatives to random assignment" (1988:70). Their approach to analysis was a step-by-step breakdown from the assignment of cases to understanding the reasons for violations. This allowed them to identify a pattern of re-assignments of more troublesome offenders who had been originally assigned to the advice or separation treatments. The effect was to bias the sample against the main finding of the RCT that arrests deterred future assaults. Their analysis also suggested the need to design similar experiments with a mechanism to allow overt re-assignment where the officer's discretionary judgment suggested that the risk posed by the offender provided a justification for over-riding the treatment protocol.

Indeed, it is quite clear that field RCTs in policing have an operational and professional context that make “occasional alternatives to random assignment” both inevitable and, from time to time, operationally required (Feder et al., 2000). Dennis (1988) divided these potential violations of the random assignment process into “overt” and “covert”: the former were planned for and mechanisms provided in the research design; the latter were both unlooked for and undesirable. If the overt and covert violations are added to the occasions where treatment is not delivered according to the agreed protocol, the question arises as to whether it is possible to allow for the potential bias in the sample that this creates and if so, how this can be done?

In considering this dilemma, Shadish, Cook and Campbell (2002) distinguished between “efficacy” and “effectiveness” trials. In a similar way, Sherman and Strang (2004) separated testing theory from testing policy and practice. In field experiments in policing, the focus is on testing the relative effectiveness of the treatment under real world conditions. As such “full treatment implementation is not always necessary or desirable” (Shadish, Cook and Campbell, 2002:319), because the trial should be seeking to find out how well the treatment will perform in the less than ideal circumstances of the operational environment. If we accept this proposition, then two linked questions arise. Firstly, what is the most appropriate balance between tight treatment integrity and real-world operational conditions? Secondly, how far is it possible and proper to compensate for potential bias’s that will inevitably result from the latter?

Shadish, Cook and Campbell (2002) suggest that there are three approaches to the analysis of treatment implementation data from an experiment: intention-to-treat analysis (ITT); analysis by the amount of treatment received; instrumental variable analysis. Weisburd, Petrosino and Mason (1993) added a fourth consideration to all three – the question as to whether increasing the sample size can be used to overcome problems with the power of an experiment.

2.2. (a) Intention to Treat Analysis

There are benefits and issues connected to all three approaches, but the consensus lies in favour of using ITT as the default approach (Lee et al., 1991). Torgerson and Torgerson (2008) demonstrated the key reason for this assumption is that any alternative approach, such as removing those who did not receive treatment or analysing the sample by treatment received rather than treatment assigned, is likely to have the effect of biasing the sample. Only an ITT approach preserves the advantages of random assignment. However, Shadish, Cook and Campbell (2002) and Torgerson and Torgerson (2008) acknowledge that ITT can be problematic when there is a high level of attrition from the trial and "random missing outcome data" (Shadish, Cook and Campbell, 2002: 320).

2.2. (b) Treatment as received

Nevertheless, the alternative of relying on analysis by treatment received is an explicit admission of the failure of the RCT design. It also runs up against significant problems in determining which subjects should be regarded as having been treated and which not. Even in the highly defined environment of an experiment to test a medicine, Lee et al. (1991) found that an ITT approach was significantly less biased than a treatment analysis because the latter both reduced the sample size and statistical power and suffered from problems with the operational definition of treatment.

2.2. (c) Instrumental Variable Analysis

Torgerson and Torgerson (2008) describe the third potential approach – instrumental variable analysis – as "in principle" "unbiased" (p.144). Angrist et al. (1996) proposed this approach and subsequently (Angrist, 2006) recommended it as a means of solving the "major statistical problems that arise in imperfect criminological experiments" (p.23). Angrist et al. (1996) started from the basic proposition that the participants in a clinical trial can be

divided into four broad groups: “always-takers” who consistently comply with the treatment; never-takers; compliers, who comply with whatever treatment they are assigned to; defiers, who do the opposite of what they are assigned to. Angrist’s IV method assumes that one or more variables can be identified in the experiment that are unrelated to the causal processes and which can be used to estimate the difference between the result from delivery or non-delivery of the treatment. In Torgerson and Torgerson’s (2008) IV example, the level of likely non-compliers in treatment and control sample was used as the variable. When Angrist (1996) applied IV to the Minneapolis Domestic Violence Experiment he used the two endogenous variables of “advise” and “separate”. He concluded that the IV analysis suggested that arrest had a stronger deterrent effect than the original analysis in Sherman and Berk (1984).

However, IV, like ITT, has some limitations, which were highlighted by Heitjan (1999) who used ITT and IV methods in parallel to assess the same dataset. He found that IV methods had the same problem with subjectivity as the analysis was based on a judgment of the extent to which subjects could be recorded “as treated”. In ITT, the researcher needed to make judgments about the extent of treatment, in IV analysis it was about the extent of estimated compliance or non-compliance. Heitjan concluded that the best evaluations should, as far as possible, rely on the simplest analysis such as a *t* test. Non-compliance leading to flaws in the level of treatment integrity meant that the simplest and best methods had to be supplemented by methods such as IV, but they were not a substitute for securing high levels of treatment integrity in the first place.

2.2. (d) Sample size and statistical power

Weisburd et al. (1993) explored one further aspect of treatment integrity – sample size and its relationship with the statistical power of the experiment. They analysed 76 criminal justice experiments and compared the effect sizes in experiments with small, moderate and large sample sizes. Contrary to the

natural intuition that effect sizes should be larger in the last category, they found, instead, that the largest effect sizes were to be found in the smallest experiments. In the larger experiments the apparent benefits of a larger sample were outweighed by implementation problems and the greater heterogeneity of the sample. “Weisburd’s paradox” suggests that “there will be few gains from increasing sample size until the design difficulties that larger sample sizes pose are addressed” (Weisburd et al, 1993: 367).

2.2. (e) Summary: the importance of treatment integrity

The justification for conducting randomised controlled trials is based on the ability of the design to filter out the noise of an operationally complex social environment and allow sound and unbiased conclusions to be drawn about the effectiveness of an intervention. However, as this overview of the literature on RCTs has demonstrated, the method requires skillful implementation and an attention to tight discipline in conduct and management. Without this, as Clarke and Cornish (1972) argued and Dennis (1988 and 1990) demonstrated, there are serious threats to the internal validity of the experiment.

It is clear that paying attention to randomisation and treatment is critical, because it is more difficult to overcome problems with these aspects of an experiment than others. Whilst there are a number of approaches to analysis that can reduce the impact of bias where random assignment has been violated or treatments have not been delivered or accepted, they cannot wholly compensate. Expanding the trial to get a larger sample does not provide a solution, nor even does keeping it small. As Berk, Smyth and Sherman (1988) and Weisburd, Petrosino (1993) argued, a tight discipline in conducting and managing RCTs is essential to prevent problems and reduce threats to treatment integrity. The question that this raises is the extent to which the completed police RCTs have been conducted and managed in this way.

2.3 Searching for RCTs in Policing:

Although there is no single recognized and comprehensive register of police experiments, there has been an increasing interest in “what works?” in policing and in systematic reviews of policing interventions. This interest has grown considerably since 2010. Notable developments have included: the George Mason Center for Evidence-Based Crime Policy “Evidence-Based Policing Matrix” (Lum et al., 2011); the UK College of Policing “What Works Centre” (NCP, 2016); the “Global Policing Database” (Higginson et al., 2015). There has also been substantial increase in the number of completed policing systematic reviews as a result of funding from the National Policing Improvement Agency (Gill et al., 2015). Yet none of these provides a comprehensive list of all Police RCTs: the GMU Matrix includes only those RCTs that bear on the police contribution to crime prevention; the What Works Centre has pursued an intervention based approach, which is, to date, far from comprehensive; the Global Police Database is, at time of writing, only fully searchable back to 2014, but promises to make the search for a comprehensive register of police experiments less labour intensive.

There have, however, been a number of important published searches for and analyses of policing and criminal justice RCTs over the last thirty years. The following have been identified as critical for this study: Dennis (1988); Weisburd, Sherman and Petrosino (1990); Petrosino (1997); Sherman et al. (1997); Farrington and Welsh (2006); Lum et al. (2011); Braga et al. (2014). A key starting point for the search strategy for this study was a review of these publications, the studies that they listed and the search criteria that they adopted. Table 2.1 shows the number of police RCTs presented in each one of these studies:

Authors	**Number of RCTs**
Dennis (1988)	17
Weisburd, Sherman & Petrosino (1990)	11
Petrosino (1997)	9

Sherman et al. (1997)	16
Farrington and Welsh (2006)	17
Lum et al. (2011) (updated 2014)	44
Braga et al. (2014)	63

Table 2.1: Reviews of Crime and Justice Randomised Controlled Trials and the numbers of Police RCTs listed.

Dennis (1988) provided the most comprehensive search and review of RCTs from 1957-1987. Dennis restricted his detailed analysis to 41 US-based studies completed between 1973 and 1987, but listed more than 300 studies dating back to 1957, including 17 RCTs involving the police. His search parameters were set to include all randomised controlled trials in criminal and civil justice research. After reviewing existing bibliographies, his principal source for additional studies was an electronic search of the National Criminal Justice Reference Service, supplemented by making contact with funding and research organisations to request unpublished technical reports and papers. Dennis discovered that he had to repeat his search several times with different search terms, because so many varied terms had been used to describe RCTs. Perry et al. (2010) have since commented on the problems created for researchers by the absence of a standard approach to the reporting and recording of RCTs.

Weisburd, Sherman and Petrosino (1990) published a “Registry of Randomized Criminal Justice Experiments in Sanctions”. The criteria for inclusion were centred on RCTs that tested a government agency’s use of a coercive condition or sanction in response to crime or to prevent crime. Studies with a sample size of less than 15 for one or more of the treatment groups were excluded. The Registry included 11 RCTs in which the police were involved. Petrosino (1997) subsequently completed a meta-analysis of randomized field experiments, which involved individual-level interventions designed to reduce reoffending. This more narrowly focused study included 9 police involved RCTs. Weisburd, Sherman and Petrosino (1990) and

Petrosino (1997) used a combination of structured electronic searching, hand-searching of journals in the field and searching through the bibliographies in pre-existing reviews and meta-analyses (Petrosino, 1995).

None of the searches reviewed above were, however, specific to policing. Sherman's chapter on "Policing for Crime Prevention" in the wider review of "Preventing Crime" (Sherman et al., 1997) was entirely focused on policing, but only included crime prevention programmes. The searches for the review primarily involved US-based evaluations of locally and nationally funded crime prevention programmes. Sherman's chapter listed 16 police RCTs.

Farrington and Welsh (2004) were building on Farrington's earlier (1983) study of randomised controlled trials in "crime and justice". Farrington (1983) had searched and found 37 studies. Farrington and Welsh (2006) found a further 85. Their criteria excluded studies with a total sample of less than 100 and included only those in which the outcome measure included a measure of offending. They included Binder and Newkirk (1977) even though their published material is far from clear on the exact sample size. Farrington and Welsh (2006) relied on extracting the studies from previous reviews, key journals, NCJRS and the SPECTR (Petrosino et al., 2000) databases and contact with key authors in the field. They specifically separated out the RCTs from each phase of the criminal justice process and included 17 RCTs, which they characterized as specific to policing.

Lum, Koper and Telep (2011) described the development of the Evidence-Based Policing Matrix. The aim of the Matrix was to provide scholars and practitioners with a readily accessible and regularly updated resource that would provide comprehensive coverage of police initiated and police involved crime prevention interventions. The Matrix has already been updated a number of times since the original searches in 2009. The eligibility criteria included only studies that met a "moderate" level on the Maryland Scale of scientific methods (Sherman et al., 1997) and, therefore, the Matrix has quasi-experiments alongside RCTs. The December 2014 update included 44 police RCTs amongst the 144 studies in the spreadsheet. The searching approach

used existing reviews of police literature, systematic reviews on policing, the National Research Council Report on Fairness and Effectiveness in Policing (Skogan and Frydl, 2004), library databases, government agencies and foundations.

Finally, Braga et al. (2014) have conducted the most comprehensive search for police RCTs of all the reviews discussed here. Their purpose was to study policing experiments in general and the process of their production together with the professional and funding network, in particular. They defined a policing experiment as one that "involved police departments in a central programmatic role". The scope of policing experiments was not restricted, as in the other searches, to experiments concerned with crime reduction, criminal sanctions or crime prevention, but included any experiment, apart from laboratory-based experiments, involving the police.

Their eligibility criteria were set by reference to Cook and Campbell's (1979) standards for a "randomized controlled design". As a result, they excluded experiments such as the Kansas City Preventive Patrol Experiment (Kelling et al., 1974) in which an intended random allocation had not subsequently been implemented. On the other hand, they included experiments, such as the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984), which had "modest implementation problems" or violations of the random assignment process.

Braga et al. described their searching strategy as "exhaustive" (2014:8). It started with a keyword search of 17 online abstract databases and included reviewing the bibliographies of previous reviews both on policing and criminal justice, the bibliographies of policing experiments, the bibliographies of systematic reviews of policing interventions and hand-searching of key journals. This combined approach produced nearly 17,000 abstracts which matched their search terms: "randomized controlled trial AND police, randomized experiment AND police and experiment AND police" (p. 8). These abstracts were then shortlisted down to 355, which appeared to show evidence of an RCT, before a draft list was developed. As a final step, in order

to help find any missing studies, the draft list of RCTs was shared with the authors identified in the studies, the authors involved with the National Academy Review (Skogan and Frydl, 2004) and other leading scholars in the field. The search was designed to cover all studies completed in 2011 or earlier and found 63 RCTs completed between 1970 and 2011.

However, even with such an "exhaustive" strategy, the Braga et al. search did not find some earlier studies, which would appear to be eligible against their screening criteria. Firstly, they do not appear, from their bibliography, to have had access to Dennis' PhD dissertation (1988), which was only available in print, rather than electronic form, and was provided personally to this researcher by the author as a scanned copy. Dennis' search back to the 1950's contained a number of early RCTs, such as Earle (1973), Ku and Blew (1977) and Lincoln et al. (1977), which are not included in Braga et al.'s list. As a result, Braga et al.'s analysis of the growth of police RCTs (at Figure 1, page 8) appears to need updating and we shall return to this below.

Secondly, the reporting and abstracting of some reports may have allowed them to escape 'underneath the wire' of the search terms that Braga et al. (2014) used. There were three RCTs designed to test "Differential Police Response" which were funded by the National Institute of Justice (All three are reported in McEwen, Connors and Cohen, 1986). The NIJ paper was entitled "an Evaluation of the Differential Police Response Field Test" and it is only by reading through the detailed methodology of the 277-page report that the RCT method becomes apparent. It does not appear in the abstracts or title. In a similar vein, Dunford, Osgood and Weichelsbaum's (1982) "National Evaluation of Diversion Pilots" contains neither the word "Police" nor any of the other search terms in the title, but yet contains four pre-court, police involved diversion RCTs as well as reporting a number of others that were abandoned due to implementation problems.

Dunford et al. (1982) could possibly have been considered for exclusion by Braga et al. (2014) because all four RCTs report significant implementation problems. Indeed, Earle (1973), one of the earliest police RCTs was

implemented so badly in the first phase that Earle and his team started again and repeated the experiment with a new sample after removing the main problems in the first phase. However, for purposes of this study, in which we are concerned with the treatment integrity of police RCTs, it is important to include rather than exclude studies which were implemented in the field as randomized controlled trials, even though they encountered more than "modest" implementation problems.

The criteria for this study have, therefore, been set more widely than any of the studies discussed above. All studies that involved the implementation of a randomised controlled design and involved the police as a significant partner in the intervention tested have been included. As with Braga et al. (2014) purely laboratory experiments have been excluded, because the emphasis of the study is on analyzing tests that were carried out in an operational, real-world environment. Braga et al. identified 54 reports and article, which described 63 eligible studies. Following a similar approach, where reports or articles, such as Dunford et al. (1982), provide details of 4 studies carried out at separate sites and reports them as separate experiments with individual data for treatment integrity, these have been treated as separate RCTs in this research.

Given the analysis set out by Weisburd et al. (2003) and Holman et al. (2016) that the size of the sample, whether large or small, is no guarantee of high levels of treatment integrity, the only exclusion for sample size has been to remove RCTs with less than 20 units in the sample where the unit of analysis is based on individuals and less than 10 units where the unit of analysis is a place or hotspot. Finally, all police RCTs completed and reported by the end of 2015 have been included. The criteria for being "reported" includes any police RCT which has been published or reported at a conference or for which the author has been able to obtain a report with sufficient detail to reliably understand the treatment integrity.

The method of searching has drawn on those described in the past narrative and systematic searches. Given that the scale of effort required to replicate

Braga et al.'s (2014) "exhaustive" systematic search was beyond the resources of the researcher, reliance has been placed on their search as a key building block. Their results were then compared with the lists from Dennis (1988), Sherman, Weisburd and Petrosino (1990), Farrington and Welsh (2004), Petrosino (1997), Sherman et al. (1997) and Lum et al. (2011) (see Table 2.1 and Appendix 1). This produced a composite list of 66 police RCTs completed prior to 2012.

In order to cover the period from the end of Braga et al.'s search (the end of 2011) until the beginning of 2016, a combination of methods was used:

- The journals searched by Braga et al. for the period 1970-2011 were re-searched for the period 2012-2016.
- The Journal of Experimental Criminology, which was not in Braga et al.'s list of Journals, was searched from its inception to the end of 2015.
- The websites of key government agencies and foundations that Braga et al. documented as funders of RCTs have been searched (the list was provided at p.11 of Braga et al., 2014).
- The "Grey Literature Database" at Rutgers University (www. http://njlaw.rutgers.edu/cj/gray) was searched using Braga et al.'s search terms together with the terms "Evaluation AND police".
- The panels and poster sessions of a number of key international conferences were searched. These included the all the Annual meetings of the American, British and European Societies of Criminology, the Society of Evidence-based Policing conference, the Cambridge Evidenced-based Policing Conferences, the Scottish Institute for Policing Research Conferences, the Jerry Lee Seminars, Campbell Colloquia, Annual CEPOL (European Police College) Seminars, the Stockholm Criminology Symposia and the George Mason Seminars.
- The bibliographies of Campbell Systematic reviews completed since 2011 were searched.

- Regular searches were made using Google Scholar and Researchgate.net, together with the NCJRS, NIJ, College of Policing and AIPM websites. In addition, alerts were set on ZETOC and Google.
- As the list of RCTs increased, the bibliographies of all the identified RCTs were searched.

Table 2.2 below shows a list of the Journals and Websites where articles and reports of police RCTs have been located both by previous reviews and by this research.

Journal, Organisation or website source	**No. of RCTs**
Journal of Experimental Criminology	16
Justice Quarterly	10
Criminology	8
Police Foundation	6
Gov.uk (for UK Ministry of Justice)	5
Aic.gov.au	4
US DoJ Office for Juvenile Justice and Delinquency	4
National College of Policing	3
Journal of Criminal Law and Criminology	3
Criminal Justice and Behavior	2
Crime and Delinquency	2
University of Queensland	2
National Institute of Justice	2
American Sociological Review	2
Cambridge Institute of Criminology Library	2
Pipersville, Pipers Press	2
Social Work Research and Abstracts	2
International Journal of Stress Management	2
Canadian Journal of Criminology	1
Preventive Medicine	1

Crime Prevention Studies	1
Translational Criminology	1
Police Quarterly	1
Yale University	1
NBER	1
Vera Institute	1
University of North Carolina	1
National Institute of Law Enforcement and Criminal Justice	1
Law and Human Behavior	1
Health Education and Behavior	1
Crime Prevention Review	1
American Journal of Police	1
Evaluation Review	1
Archives of Pediatrics and Adolescent Medicine	1
Health Education and Research	1
British Journal of Criminology	1
Journal of Research in Crime and Delinquency	1
Drug Alcohol and Dependence	1
Juvenile Justice	1
Urban Institute	1
NCJRS abstracts	1
National Science Foundation of China	1
Scottish Institute for Policing Research	1
Fundacion Ideas para la Paz	1
Static1.1.sqspcdn.com	1
Rutgers University	1
Amazon.com	1
University of Maryland	1
Journal of Applied Social Psychology	1
TOTAL	**122**

Table 2.2: Journal, Organisation and Website sources for Police RCTs

These combined methods produced a draft list of 107 completed police RCTs and some 30 “in-flight” RCTs which was then circulated to the some of the key police RCT researchers in Australia, USA, Israel, UK and Europe, in the same way as Braga et al., in order to quality assure the list and identify any missing studies. This, together with a set of studies completed and published in 2015, has produced a final list of 122 RCTs. Thirty-two of these RCTs were completed and reported after the end date of Braga et al.’s (2014) search.

Although the search has found almost double the number of police RCTs compared to the Braga et al. (2014) search and appears to provide the most comprehensive list of police RCTs yet collected, it is important to record the limitations to this search. Although the aim has been to be as comprehensive as possible, using methods similar to Farrington and Welsh (2004) and Lum et al. (2011), it was a not a systematic search as set out in the Cochrane or Campbell methods (Higgins and Green, 2008 and Campbell Collaboration, 2014). The scope and eligibility criteria for this search have been much broader than those set by any of the earlier searches. The intention has been to find as many completed RCTs involving the police as possible without the dedicated time and resources for a full systematic search. The major reason for this is the pragmatic recognition that in contrast to a subject or intervention based systematic review where at least the scope of the search and the likely sources are to some extent limited by the topic, a search for all police RCTs on every topic and intervention would require a level of resourcing well beyond the individual researcher and timeframe for this research. Only with the completion of the Global Police Database project is such an enterprise likely to be realistic and even then, on the evidence of this search, there may yet be studies that evade the search terms.

For this reason, it is also difficult to estimate the potential publication bias. Publication bias is a significant risk in systematic reviews because, as Torgerson (2003) sets out, meta-analysis can be skewed by the absence of unpublished research from the studies analysed. In systematic reviews

researchers are recommended to use techniques such as “funnel plots” to assess the risks of publication bias. However, the focus in this research is on the problems of treatment integrity across the highly heterogeneous field of all police experiments. Techniques such as funnel plotting, which rely on comparing effect sizes from published and unpublished literature looking at similar interventions, would not seem appropriate. Even though the search has sought to include the Grey literature, the potential risk of bias from the non-publication of failed, flawed and unsuccessful experiments is a significant one that needs to borne in mind in interpreting the analysis. The author is aware from a number of personal communications with scholars to whom the draft list was circulated that there have been a number of failed, unreported police RCTs.

Torgerson (2003) and Lipsey and Wilson (2001) also identified another potential bias that is undoubtedly relevant to this search. Some studies have been published with key data missing or have been published before some types of data or key protocols (such as CONSORT) have become accepted standards (Perry et al., 2010). This is a particular problem with studying 50 years of police RCTs because the reporting requirements and methodological research knowledge has changed so significantly over that time. Furthermore, not only have publication standards changed over time, but also some of the reports located were published for an operational audience rather than peer-reviewed publication. As a result, the data presented in Appendix 1 and analysed in more detail below contains some significant gaps. Lipsey and Wilson (2001) nevertheless encouraged the researcher to present all the studies even where there is missing data. However, the variability of the data and the range of topics and designs included by the search has meant that the analysis approach has been determined by a need to find a method that provides as reliable an estimate of treatment integrity as possible given the unevenness of the data quality.

2.4 Methods: Analysing the “Matrix of Police Experiments”

The aim of the analysis presented in this research is to assess the level of treatment integrity in completed police RCTs and then use that analysis to draw out the implications for the conduct and management of such experiments in policing. There is a substantial literature on assessing the quality of RCTs and the development of scales and checklists in order to do so (Juni, 2001). Moher et al. (1995) critically reviewed these and found wide variation in approach and significant weaknesses in the underpinning justifications for key choices. They concluded that scales and checklists were most helpful for the guidance of authors reporting their studies rather than researchers assessing their strengths and potential bias. Higgins and Altman (2008) also observed that many of the scales and items in the checklists bore little or no relevance to assessing the internal validity of the experiment.

Higgins and Altman (2008), in setting out the approach for assessing internal validity and bias in Cochrane Systematic Reviews, rejected scales and checklists in favour of a subjective domain-based evaluation. In arriving at this approach, the Cochrane methods group had concluded that it was impossible to know the full extent of bias in a study and that the "most realistic assessment of validity of a study may involve subjectivity" (8.3.1). Using scales and checklists, whilst offering a plausible and simple scoring system were demonstrably unreliable (Juni, 1999).

The Cochrane approach (Higgins and Altman, 2008) revolves, instead, around a transparent process of gathering and assessing data from the published reports and protocols of each study against five areas of bias: selection; attrition; performance; detection; reporting. For the purposes of this analysis, two of these areas of bias - attrition and performance bias – appear to be most relevant. Performance bias refers to the risk of systematic differences between groups in the level of treatment provided. Attrition bias is concerned with the level of differential attrition from the groups in the study. Taken together, these two areas concern the risks of bias that arise from problems with ensuring the allocated sample(s) are sustained and treated as intended through the experiment.

In the Cochrane model the risk from the five areas of bias is assessed by using the criteria in the Cochrane Collaboration "risk of bias" assessment tool (Higgins and Altman, 2008). Whilst this process is well designed for researchers weighing the risk of bias in studies being considered for inclusion in a systematic review of an intervention, the requirement for this study is more focused. On the other hand, the approach that Cochrane recommends for forming and expressing the judgment on bias has much to recommend it. Higgins and Altman (2008) set out a simple banding of low, unclear and high risk supported by an evidence trail showing the rationale for the judgment, which is then set out in a "risk of bias summary". Applied to the sample of 122 completed police RCTs, a version of this approach would appear to offer a potentially realistic, evidenced and justifiable way of differentiating between low, medium and high treatment integrity experiments.

Drawing, therefore, on Higgins and Altman (2008) and the structured, subjective assessment of data from RCTs, the approach taken to assessing each of the completed police RCTs identified in the search has used a three step process: building a report card for each reported experiment from the available articles and publications: creating a "Matrix" of the experiments recording key details (Appendix 1); estimating the treatment integrity of each experiment from the available data supported by criteria set out in Table 2.3.

2.4. (a) Report Card:

Dennis (1988) adopted a "report card" approach in assessing the internal validity of criminal justice RCTs. Weisburd, Sherman and Petrosino (1990) present a narrative report card on each of their listed experiments. A similar approach has been adopted in this study. Starting with the articles and reports, the data from each has been developed using two approaches: building a CONSORT flow diagram for each study; noting key information about the process of assignment and treatment in order to support the judgment against the criteria in Table 2.3. The CONSORT diagrams used the outline provided by Perry et al. (2010).

2.4. (b) Matrix of Experiments:

A “Matrix of Police Experiments” was created with details of all the studies. This drew on the design of the Evidence-based Policing Matrix (Lum et al., 2011). The Matrix recorded the following fields of information: the publication references; key hypotheses; coding by study topic (for which Braga et al.’s (2014) coding has been used as a basis); study outcomes and summary; coding for the significance, non-significance or back-fire effect of the study; coding for whether the study was identified in previous searches (as per Table 2.1 above); sample size and unit of analysis.

Appendix 1 presents an edited summary of key information from the Matrix, which supports the analysis below: author reference; year of publication; topic and coding; sample size and unit of sample; treatment integrity estimate; three columns, which we will discuss below in Chapter 3 which present an analysis of whether the research team were police RCT novices, whether the RCT was conducted in a new research “station” and whether it was a novel topic for a police RCT.

2.4. (c) Grading the experiments:

The final stage was to estimate the treatment integrity of each study. Higgins and Altman (2008) suggest a structured assessment against a set of criteria. Table 2.3 sets out the key criteria which were considered in assessing the treatment integrity of the 122 RCTs. The criteria have been divided in to Attrition and Treatment Performance, reflecting the Cochrane model for doing so.

Definitions	**Key areas considered in the analysis**	**Data thresholds for GREEN, AMBER, RED and UNCLEAR**

Attrition: The extent to which the selected sample was sustained through the trial and consistently treated as assigned	• Is there a CONSORT diagram? • How is the Random assignment process implemented? • Is there clear evidence of its oversight and control? • Is there evidence of tracking and monitoring? • Is there a process for overt override and data provided?	90% + Low risk of bias (GREEN) 60% + medium risk of bias (AMBER) -60% High Risk of Bias (RED) Unclear Insufficient data to make a clear judgment
Treatment Performance: The extent to which the treatment required to be delivered in order to test the hypothesis was delivered as intended and in accordance with the treatment protocol	• Is there a CONSORT diagram? • Is there is a clear narrative on the implementation of treatment and the dosage delivered? • Is there Evidence of tracking and monitoring? • Is there a treatment protocol or clear narrative on treatment requirements?	90% + Low risk of bias (GREEN) 60% + medium risk of bias (AMBER) -60% High Risk of Bias (RED) Unclear Insufficient data to make a clear judgment

Table 2.3: Key criteria for Attrition and Treatment delivery estimates:

The criteria were used by the researcher in structuring the report cards for each RCT and noting key points on the conduct and management of the experiment. The overall figure for treatment integrity was calculated by using the data in the CONSORT constructed for each police RCT. It provides an aggregate estimate of the combined effect of the percentage of cases treated as assigned and to the intended treatment dosage. In all three assessments, as suggested by Higgins and Altman (2008) it has also been necessary to provide an "unclear" estimate where the key data is missing from the study.

The Cochrane "risk of bias", whilst providing a useful starting point for this analysis, provides only a binary judgment for risk of bias. Each study is required to be assessed as either "low risk" or not, with an "unclear" category for studies with missing data. However, the purpose of the Cochrane analysis is to triage out studies, which do not have a low risk of bias. In this research, we are not concerned with the omission of studies, because all the studies in the Matrix have met the eligibility criteria for inclusion. Hence, there has been a need to adapt the Cochrane approach and find a means of estimating whether each RCT has a high (GREEN), low (RED) or "in-between" (AMBER) level of treatment integrity.

As we have seen in the discussion of treatment integrity above, there is an acceptance that in a field trial in the "real world" there is likely to be some level of slippage to treatment integrity (Cook and Campbell, 1979). Boruch (1997) suggested that a 10-15% rate of "discrepancy" between "treatments assigned and those delivered is high enough to warrant investigation" (p.106). Boruch based this estimate on his own and Dennis' (1988) work on the implementation of RCTs. Given the basis for Boruch's judgment was formed from a wide-ranging review of RCTs from education, social welfare and criminal justice (supported by Dennis' comprehensive analysis), it seems reasonable to use his suggestion as a guide for the threshold at which to set an estimate of "high" treatment integrity, which is, therefore, set at 90% or above.

In order to set a reasonable and justifiable threshold for the boundary between low and “in between” or medium, we have drawn on the literature on program implementation. Durlak and Dupre (2008) carried out a meta-analysis of the factors affecting implementation in youth, prevention and health promotion programmes. They found that there was a positive correlation between effect sizes and levels of implementation. They observed that “expecting perfect or near perfect implementation is unrealistic. Positive results have often been obtained with levels around 60%; few studies have attained levels greater than 80%. No study has documented 100% implementation for all providers” (p.331). This finding was based on assessments of the implementation of programmes or interventions rather than simply the level of treatment integrity in experimental conditions. However, Durlak and Dupre (2008) identified that the delivery of the intended dosage and the monitoring of control and treatment conditions were two of the key aspects of implementation (out of eight they documented). This suggests that it is credible to deploy their 60% threshold, below which they imply that credible results are less likely to be obtained, as the division between low and medium treatment integrity.

The “unclear” category has been applied when the articles or reports for the RCT have either failed to provide key data or where the data provided is ambiguous. Some caution needs to be exercised in interpreting the patterns of “unclear” over time, because, as Perry et al. (2010) suggest, standards of reporting and the expectations for journal publication have changed over time. The publication of standards by the Society for Prevention Research (Flay et al., 2005 and Gottfredson et al., 2015), in particular, have raised the bar for data requirements.

In summary, drawing on Boruch (1997) and Durlak and Dupre (2008), we have set the thresholds for “high”, “medium” and “low” treatment integrity at 90%, above 60% and below 60%, with the addition of an “unclear” category where the data is missing or ambiguous.

2.5 Findings

This section provides an analysis of demographics and treatment integrity of the 122 experiments: Table 2.4 provides an overview of the analysis, which is drawn from the full Matrix presented at Appendix 1:

- The experiments, their author and the dates completed;
- The types of subjects studied – the codes draw on and expand those used by Braga et al (2014) in their analysis;
- The total sample sizes and unit of analysis;
- Four-colour banded scores for the level of overall treatment integrity.

Author(s)	Year	Coded RCT topic ((after Braga et al., 2014)	SAMPLE	UNIT	Treatment integrity %
Abrahamse et al.	1991	offender management	480	Offenders	91.25
Ackerley	1986	Health and Welfare	49	Police Officers	U
Amendola et al.	2011	Health and Welfare	326	Police Officers	84.35
Angel et al.	2014	Restorative Justice	192	Victims	99.47
Antrobus	2015	Crime Victim outreach	978	Crime Scenes	93.5
Ariel et al.	2014	Body Worn Cameras	988	Shifts/Tours of duty	U
Banerjee et al.	2013	Road Safety - Drink Drive	123	Roadside Breath Stations	53
Banerjee et al.	2012	Administrative Reforms	162	Police Stations	84
Bennett & Newman	2015	Fear Reduction	26	Hotspots in N Brisbane	U
Berk et al.	1992	Domestic violence	1658	DV suspects	82
Binder & Newkirk	1977	Juvenile justice	U	Juvenile offenders	U
Boyanowsky & Griffiths	1982	Legitimacy	133	Citizens stopped at roadside	U
Braga & Bond	2008	Hotspots/crime places	34	Hotspots	U
Braga et al.	1999	Hotspots/crime places	24	Violent Crime Hotspots	U
Byles & Maurice	1979	Juvenile justice	305	Juveniles with 2 or more priors	45
Clayton et al.	1996	DARE	31	Schools and Pupils in 6th Grade	93
Davidson et al.	1977	Juvenile justice	37	Juvenile offenders	U
Davidson et al.	1977	Juvenile justice	36	Juvenile offenders	U

Davis & Taylor	1997	Domestic violence	435	Households in which family violence occurred	84
Davis & Medina-Ariza	2001	Domestic violence	60	Housing Projects/Households	50
Davis & Maxwell	2002	Domestic violence	197	Victims of Family violence	U
Davis et al.	2007	Domestic violence	300	Victims of Domestic Violence	85
Dunford et al.	1990	Domestic violence	330	Offenders in DV cases	92
Dunford	1990	Domestic violence	247	Offenders in DV cases	96.5
Dunford et al.	1982a	Juvenile justice	433	Juvenile offenders	84.7
Dunford et al.	1982b	Juvenile justice	686	Juvenile offenders	69.5
Dunford et al.	1982c	Juvenile justice	533	Juvenile offenders	78.2
Dunford et al.	1982d	Juvenile justice	975	Juvenile offenders	54.35
Earle	1973	Police Training	174	Deputy Sheriffs	57.5
Eck & Wartell	1998	Hotspots/crime places	121	Residential properties subject of drugs enforcememt	83
Esbenson et al.	2012	DARE	195	Classrooms	90
Gersons et al.	2000	Health and Welfare	42	Police Officers	97.6
Giblin	2002	Juvenile justice	190	Juvenile offenders	45
Glick et al. [1]	1986	shoplifting	1346	Shoplifters	92
Graziano et al.	2014	Citizen feedback/interaction	51	Beats in Chicago	47.5
Groff et al.	2005	Citizen feedback/interaction	314	Residents	98
Groff et al.	2015	Hotspots/crime places	81	Micro places	U
Grossmith et al.	2015	Body Worn Cameras	1510	Officers	U
Hegarty et al.	2014	Hotspots/crime places	48	Hotspots of crime	88
Hirschel et al.	1990	Domestic violence	686	Cases of Domestic Violence	83.5
Ireland et al.	2007	Health and Welfare	129	Police Officers	51.9

[1] Sherman (personal communication, June 2017) has suggested in his final comments on this dissertation that this experiment had “unknown levels of cheating” because the control of eligibility lay with the store detectives, and, therefore, some caution needs to be exercised on the estimation.

Jennings et al.	2015	Body Worn Cameras	89	Police Officers	U
Jolin et al.	1998	Domestic violence	927	DV Incidents	70.7
Komro et al.	2004	DARE	24	Schools and 7th Grade pupils	98
Koper et al.	2013	Hotspots/crime places	117	Hot route sites	100
Ku & Blew	1977	Juvenile justice	36	Adolescent offenders	100
La Vigne & Lowry	2011	CCTV	50	Parking facilities	100
Langley	2014	Legitimacy	781	Passengers	98.4
Lincoln et al.	1977	Juvenile justice	306	Juvenile offenders	U
Little et al.	2004	Juvenile justice	90	Juvenile offenders	64
Lu et al.	2012	Road Safety: traffic enforcement	80377	Private car owners	100
Lum et al.	2011	Hotspots/crime places	30	Hotspots	84.6
Lurigio & Rosenbaum	1992	Crime Victim outreach	122	Police Recruits	26.5
MacQueen & Bradford	2014	Legitimacy	20	Police Units conducting Road Safety tests	U
Martin & Sherman	1986	Offender management	414	Repeat Offenders	80
McCold & Wachtel	1998	Restorative Justice	111	juvenile offenders	31.5
McCold & Wachtel	1998	Restorative Justice	181	juvenile offenders	49.5
McCraty & Tomasino	1999	Health and Welfare	65	Police Officers	90.7
McEwen et al.	1986a	Differential Police Response	5510	Calls for Service	92.1
McEwen et al.	1986b	Differential Police Response	34795	Calls for Service	91.3
McEwen et al.	1986c	Differential Police Response	5497	Calls for Service	98
McGarrell & Kroovand	2007	Juvenile justice	782	Juvenile offenders	98.5
Mazerolle et al.	2012	Legitimacy	60	Random Breath Test Stations	100
Mazerolle et al.	2000	Hotspots/crime places	100	Street Blocks	70
Mejia et al.	2013	Police Training	104	Police stations	U
Mohler et al.	2016	Predictive Policing	510	Patrol days	U
Neyroud et al.	2015	Pre-court diversion	417	Offenders	91
Norvell & Belles	1993	Health and Welfare		Police Officers	64.4
Owens et al.	2014	Body Worn Cameras	300	Police Officers	U
Owens et al.	2015	Legitimacy	1444	Police Officers	100
Pate et al.	1985a	Citizen feedback/interaction	660	Households	U

Pate et al.	1985b	Citizen feedback/interaction	504	Households	U
Pate & Hamilton	1992	Domestic violence	907	Cases of Domestic Violence	90
Pate et al.	1991	Domestic violence	907	Victims of Domestic Violemce	61.3
Piza et al.	2015	CCTV	38	CCTV Camera areas	U
Quay & Love	1977	Juvenile justice	568	Juvenile offenders	U
Quinton	2011	Citizen feedback/interaction	7434	Citizens	100
Ratcliffe et al.	2011	Hotspots/crime places	60	Violent Crime Hotspots	U
Ridgeway et al.	2011	Gun possession	2120	Gun purchasers	68.2
Ringwalt et al.	1991	DARE	20	Schools and pupils	100
Roman et al.	2009	DNA/Crime detection	2150	Volume crime cases	99.9
Rose & Hamilton	1970	Juvenile justice	494	Young people	U
Rosenbaum et al.	1989	Crimestoppers	44	Anonymous Callers to Crimestoppers	U
Rosenbaum et al.	1994	DARE	24	Schools and pupils	U
Rosenbaum & Lawrence	2013	Legitimacy	157	Police Recruits	U
Rosenfeld et al.	2014	Hotspots/crime places	32	Violent/Firearms Crime Hotspots	U
Sahin	2014	Legitimacy	702	Drivers stoppped in speed checks	96.6
Santos and Santos	2014	Hotspots/crime places	48	Hotspots of crime	95.9
Shapland et al.	2006a	Restorative Justice	106	Offenders	92.2
Shapland et al.	2006b	Restorative Justice	186	Offenders	90.3
Shapland et al.	2006c	Restorative Justice	165	Offenders	95.1
Shapland et al.	2006d	Restorative Justice	165	Offenders	98.2
Shapland et al.	2006e	Restorative Justice	105	Offenders	95.7
Shapland et al.	2006f	Restorative Justice	105	Offenders	93.3
Sherman et al.	1989	Hotspots/crime places	500	Repeat Call addresses	U
Sherman & Weisburd	1995	Hotspots/crime places	110	Crime Hot spots	91
Sherman & Berk	1984	Domestic violence	314	Domestic Violence cases	83
Sherman & Rogan	1995	Hotspots/crime places	207	Blocks in Kansas City	97

Sherman et al.	1992	Domestic violence	1200	Domestic Violence cases	98.25
Shipley and Baranski	2002	Health and Welfare	54	Police Officers	100
Short et al.	1984	Health and Welfare	45	Police Officers	100
Strang et al.	1999a	Restorative Justice	900	Offenders	93
Strang et al.	1999b	Restorative Justice	143	Offenders	88.5
Strang et al.	1999c	Restorative Justice	249	Offenders	76.9
Strang et al.	1999d	Restorative Justice	121	offenders	86
Skogan & Wycoff	1987	Crime Victim outreach	485	Victims of personal household crimes	85
Sloboda et al.	2009	DARE	83	Schools and pupils	75
Slothower	2015	Crime Victim outreach	142	victims of crime	91
Sousa et al.	2010	TASER	64	Police Officers	100
Stratton	1975	Juvenile justice	60	juvenile offenders	U
Tanigoshi et al.	2008	Health and Welfare	60	Police Officers	85
Taylor et al.	2001	Domestic violence	376	Court Defendants	85
Taylor et al.	2011	Hotspots/crime places	83	Violent crime hotspots	U
Telep et al.	2014	Hotspots/crime places	42	Hotspots	100
Weisburd & Green	1995	Hotspots/crime places	56	Drug Hotspots	100
Weisburd et al.	2015	Hotspots/crime places	232	Police beats	100
Weisburd et al.	2011	Hotspots/crime places	110	Street Segments	78.57
Weisburd et al.	2008	Hotspots/crime places	26	Census Blocks	U
Wells et al.	2005	Citizen feedback/interaction	57	Police Officers	84
Wells et al.	2015	Eye-witness identification	497	Line-ups	100
Wheller et al.	2013	Police Training	576	Police Officers	86.1
Wilson et al.	2001	Health and Welfare	62	Police Officers	95

Table 2.4 – treatment integrity in Police RCTs

The analysis which follows starts with the demographics of the sample and then turns to the treatment integrity of the studies identified. The demographics for this research include the year by year publication dates, the authors and the topics under study.

2.5.1. Demographics of Police RCTs 1970-2015

Tables 2.5 and 2. 6 show the year by year publication date of the studies.

Table 2. 5 also shows the year by year numbers:

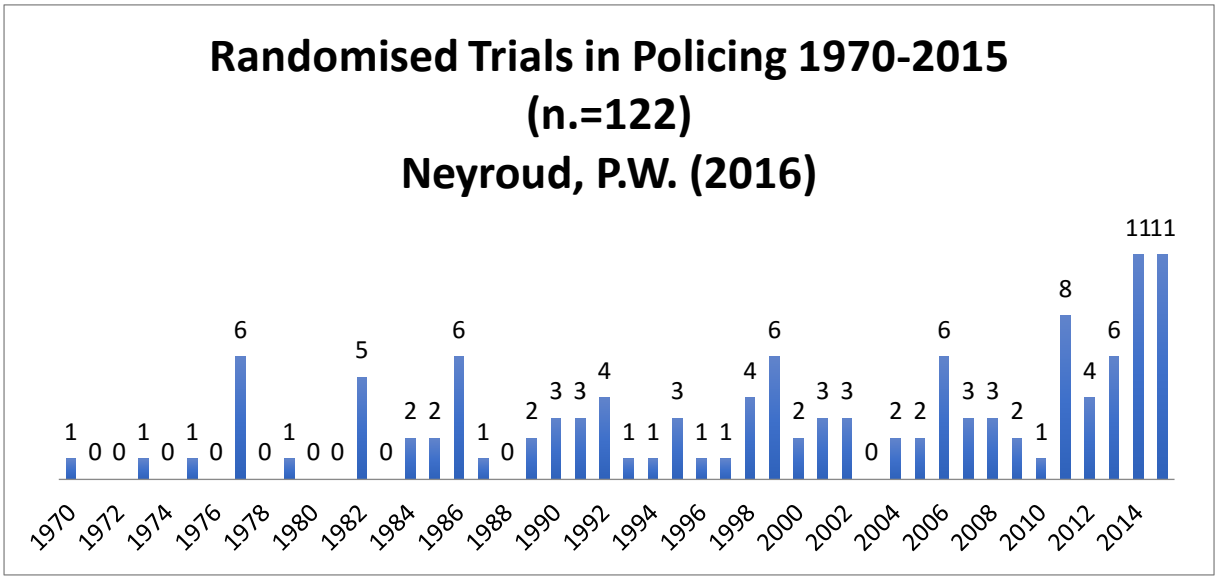


Table 2.5: Year by Year numbers of Randomised Controlled Trials in Policing 1970-2015 by year of report (n = 122)

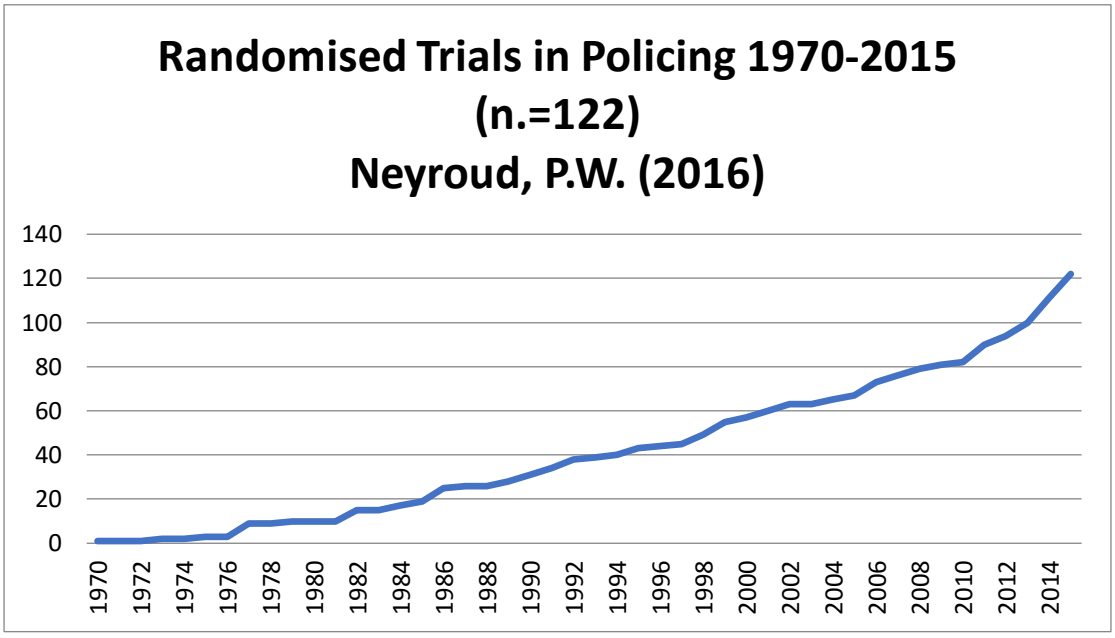


Table 2.6: Year on Year cumulative progress in numbers of Randomised Trials in Policing 1970-2015 by year of report (n=122).

Some caution needs to be exercised in interpreting the precise years to which studies are allocated, because a number of factors will have determined the date in Appendix 1 and in Table 2.4, which has been set as the date of the main report or article relating to the RCT. For instance, some RCTs have been published in government reports at the end of a grant period and will, therefore, have a publication date in relatively close proximity to the date of completion of the trial, others have been published in peer reviewed journals where the process of review and the publication queue may have imposed a significant delay. A further example is provided by Lincoln et al. (1977) and Klein (1986), who are both reporting the same study but with two different frameworks for analysis: the first reported the initial outcomes; the second a re-analysis exploring the findings against labelling theory. Braga et al. (2014) reported the Klein (1986) study, whereas Dennis (1988) reported both and linked them. The different dates for reporting the same study reflect a wider point about studies where reoffending is a key measurement criterion and the data for the outcome analysis may take twelve to twenty-four months to gather. It is, therefore, important to read Tables 2.5 and 2.6 together and to focus more on the progress across 5 years or even across each decade rather than the precise pattern of RCTs published in particular years.

With this caveat in mind, Tables 2.5 and 2.6 show a period of slow growth in the 1970's (10 RCTs 1970-1980), followed by three decades in which there was a reasonably steady pattern of 22 (1981-90), 26 (1991-2000) and 24 (2001-2010) RCTs per decade. However, in the most recent five years – 2011-2016 – there has been a dramatic increase compared to this pattern of completion with 40 RCTs in only half a decade. The number of in-flight RCTs – more than 60 are reported in Appendix 2 - would suggest that this upwards trajectory is not an isolated rise but rather a significant acceleration from the last 30 years. If the in-flight RCTs are completed and published (and given that the initiation or progress of most of them has been reported at International Conferences this seems highly likely), the decade 2011-2020 seems likely to more than match and probably surpass the total number of RCT studies completed from 1970-2011. This matches wider developments. Halpern (2017) used Shepherd's analysis of the rise in RCTs in medicine and

social science in the 20th century to describe the phenomenon as the “rise of experimental government” (Slide 1).

Within that changing trend upwards there has also been a substantial change in the researchers engaged in Police RCTs. Braga et al. (2014: Table 3) found that, in respect of their 63 studies, 126 “scholars” had been named as authors. This research has identified that that list, up to 2011, was a little longer. However, even with the addition of the extra scholars identified in the pre-2012 studies the core findings of Braga et al. remain well founded: most Police RCTs had been conducted by a “small and highly active connected network of collaborators and students” (p.21).

However, when we look at the picture between 2012 and 2015, it is clear that the rising trend in Police RCTs has been matched by an equally significant change in the scholarly network. Of the 90 named authors, only 14 were represented in Braga et al.’s list, whilst 76 (84%) were new authors. A number of the authors are from areas – South America, India and China – that were completely unrepresented prior to 2011. In just under half (14 out of the 32) of the studies there were strong connections between the new authors and one of the ten most prolific scholars from 1970-2015 (Braga et al., 2014), either as co-authors, colleagues at the same research institution or as students. However, in more than half of the studies, there were new authors who appear not to be part of the previous “connected network of collaborators and students”. Furthermore, whilst between 1970-2011 only four studies (Earle, 1973, Sherman et al., 1992, Braga et al., 1999 and Shipley and Baranski (2002)) appear to have had an author who was a serving police officer or member of staff in a policing organization, there are 11 studies from 2012-2016 with professional practitioner involvement.

Taken together, these changes – more new scholars, a wider international reach and a significant increase in professional involvement – might reasonably be interpreted as signalling the type of tipping point towards a more central role for experimentation in policing research that Braga et al. (2014) hinted at in their conclusions. An examination of the list of more than

60 in-flight studies (Appendix 2) provides further support for this hypothesis. There appear to be more “new” scholars (although without the reports or articles with a final author list it is difficult to compare definitively). Moreover, more than two thirds of the in-flight studies appear to involve practitioners as either one of the principal investigators or as a member of the research team and, therefore, likely to be included on the authors listed as and when the study is published. This suggests that there may be an increasing trend towards active practitioner involvement in the investigation process, rather than just as the commissioners, subjects or recipients of the research.

Topic coded to Braga et al. (2014) categories	**Numbers**
Hotspot/place based	20
Juvenile Justice	16
Domestic Violence	14
Restorative Justice	13
Health and Welfare	10
Legitimacy	7
DARE	6
Citizen Feedback/Interaction	6
Crime Victim Outreach	4
Body Worn Cameras	4
Police Training	3
Differential Police Response	3
Offender Management	2
Road Safety	2
CCTV	2
Pre-court Diversion	1
Fear Reduction	1
DNA in Detection	1
TASER	1
Admin Reforms	1
Shoplifting	1
Eye-Witness ID	1
Shoplifting	1
Gun Possession	1
Crimestoppers	1

Table 2.7: Police RCTs by coded category of subject studied 1970-2015 (n.122)

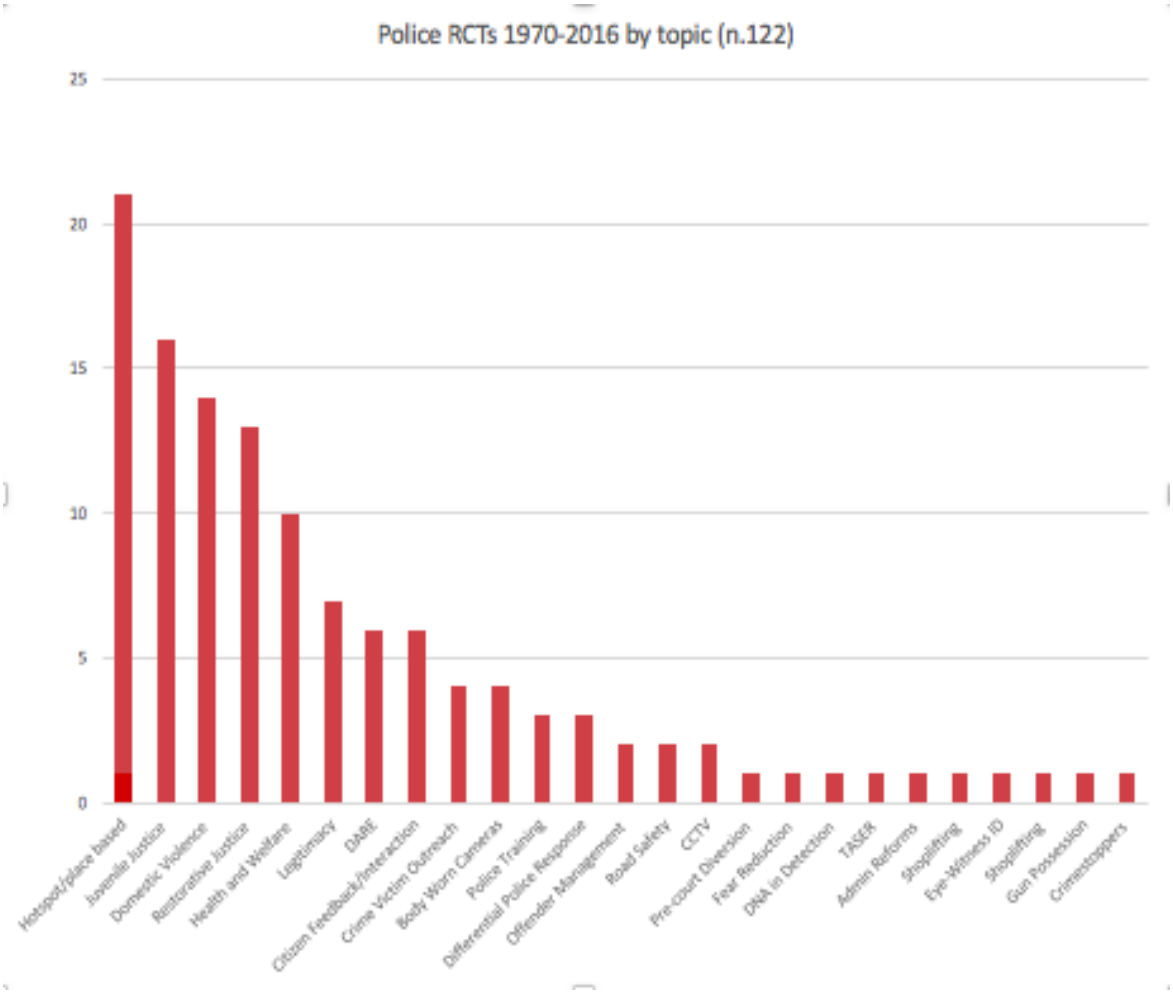


Figure 2.1: Police RCTs 1970-2016 by Topic (n.122)

This research has also identified differences between the range of topics studies as compared to Braga et al. (2014). The studies in Appendix 1 have been coded as closely as possible to the labels from Braga et al. (2014) but with the addition of a significant number of additional categories from the new studies retrieved.

Table 2.7 and Figure 2.1 show that six subjects account for more than 65% (80 out of 122 studies) of the total: hotspot/place based (20); juvenile justice (16); domestic violence (14); restorative justice (13); health and welfare (10); Legitimacy (7). In Braga et al. (2014)'s sample the first three alone (hotspots,

restorative justice and domestic violence) accounted for 65% (41 out of 63) and they listed a total of 13 subjects compared to 25 listed in Table 6. The major differences in this research are in Juvenile Justice (16 studies compared to 4), Health and Welfare (10 studies compared to 1 study on shift length), Legitimacy (7 studies compared to none), Body Worn Cameras (4 compared to none) and Differential Police Response (3 compared to none). Part of this difference can be accounted by the retrieval of a number of early studies, particularly in juvenile justice, which were listed by Dennis (1988), part also by the retrieval of studies such as McEwen et al. (1986) in which there was no reference to the RCT method in the title, abstract or searchable entry in NCJRS. Thirdly, the ten Health and Welfare studies were listed in Patterson et al.'s (2012) Campbell Systematic Review on Stress Management Interventions in policing, which was completed after Braga et al.'s search. They were also all published in journals outside Braga et al.'s search criteria and none included the search terms in their titles.

However, there is also evidence of a shift in the international research agenda - all 11 Legitimacy and Body Worn Camera studies have been completed and published since 2012, following a period of increased concern about and interest in police legitimacy (for example President's Taskforce on 21st Century Policing, 2016). That agenda has continued to be important, as evidenced by the considerable number of Body Worn Camera RCTs in flight (Lum et al., 2015 and Appendix 2).

All of this suggests a need to make some adjustment to Braga et al.'s finding that police RCTs have been concentrated in a "small number of policy areas" (p.11). This analysis suggests that the range has been a little wider and is now expanding in response to policy and practice concerns, especially about police reform and police legitimacy (Presidents Taskforce, 2015). This may also reflect the increasing engagement of practitioners not only in the conduct and management of Police RCTs, but also in networks such as the Society for Evidence-based policing, the development of research active professional bodies such as the UK College of Policing (which has conducted RCTs in both Legitimacy and Body Worn Cameras since 2012) and police-academic

partnerships such the Scottish Institute for Police Research (which was a key sponsor of MacQueen and Bradford's ScotCet legitimacy RCT (2014)).

2.5.2. Treatment integrity

The analysis of treatment integrity is set out in the right-hand column of Table 2.4. The estimates of treatment integrity are "RAG" rated (Red, Amber, Green and Unclear) and provide an estimated figure for the overall treatment integrity of the RCT calculated from the CONSORT constructed from or published in the articles or reports. The findings are set out below under the following headings: "Unclear" estimates; the overall pattern of estimates; the "red" estimates; the "green" estimates.

2.5.2 (a) "Unclear" estimates:

It is important to recognize that the estimates of treatment integrity are just that: an informed estimate based on the best evidence available to the researcher. That evidence did not always allow a clear conclusion to be drawn. There were 33 out of 122 studies (27%) where the information in the published report was insufficient to form a clear estimate on the overall treatment integrity. Comparing the first 20 years up to 1990 with the most recent decade (2006-16), the percentage of "unclear" studies has fallen from nearly half (46%) to less than a third (32%).

Twenty of the "unclear" studies are to be found in just three out of the 25 topic areas: hotspot policing; juvenile justice; body worn cameras (see Figure 2.2 below). The explanations for this vary between the three topics. In the case of the eight Hotspot studies the issue that makes it very difficult to make a clear assessment is the absence of data on the dosage of patrol in the hotspots. We know from a number of studies (for example, Koper, 1995, Telep, Mitchell and Weisburd, 2014, Santos and Santos, 2015 and Ariel, Weinborn and Sherman, 2016) that the level of dosage – the key independent variable - is important in targeted policing interventions. In assessing these studies, clarity

on levels of dosage and the tracking methodology was given careful attention as a key indicator of treatment delivery.

The juvenile justice studies with unclear data are all early studies carried out before 1980. These studies tend to lack key details in the methods section and the authors have concentrated on reporting the treatment philosophy and outcome data with insufficient attention to the conduct and management of the trial. This observation seems to be consistent with Martinson's concerns about the quality of the some of the studies reported before his review (Martinson, 1974). Given that there are also three other studies in the area of juvenile justice which appear to be below the 60% threshold (Byles and Maurice, 1979, Giblin, 2002 and Dunford et al., 1982d), the lack of detailed attention to reporting the conduct and management of these trials presents a potential concern about the risks of bias and levels of treatment integrity in the "unclear" cohort.

The final group of studies with an "unclear" assessment is four recent Body Worn Camera RCTs (Ariel et al., 2014, Grossmith et al., 2015, Owens et al., 2014 and Jennings et al., 2015). As with the hotspot studies, Ariel has observed that treatment integrity is an important issue in Body Worn Camera evaluations because there is emerging evidence of a relationship between compliance with a requirement to wear cameras and some of the key evaluated outcomes (Ariel, 2015). However, in each of these studies the researchers have either found it difficult to assess the level of implementation (Ariel et al., 2014), have provided insufficient data to make an assessment (Jennings et al., 2015) or have provided data that suggests a highly problematic implementation but without sufficient detail to estimate the treatment integrity with any confidence (Grossmith at al. 2015 and Owens et al., 2014). Both of the Owens et al. and Grossmith et al. studies, which were carried out by researchers from the College of Policing, hint at very serious implementation difficulties but without providing a clear CONSORT diagram summarizing the data on treatment integrity. Grossmith et al. commented that "while the aim was for all the officers in the treatment group to receive the intervention, in practice not all ended up being sent on training and issued

with a camera as intended" (p.10). Owens et al. lamented that "it is, therefore, impossible to tell to what extent the intervention was actually delivered" (p.11).

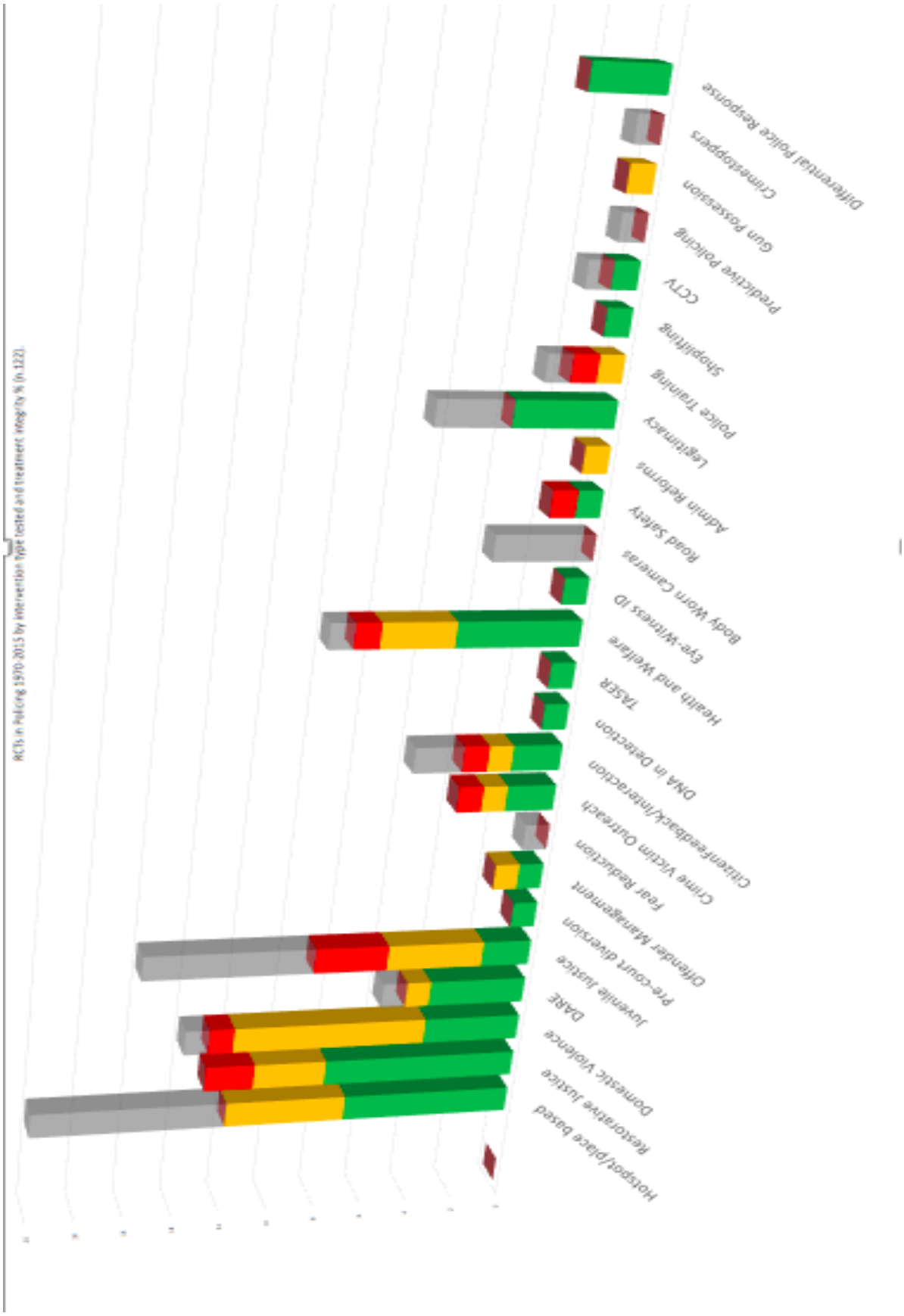


Figure 2.2: RCTs in Policing 1970-2015 by intervention type tested and treatment integrity RAGU rating (n.122)

2.5.2 (b) The overall pattern of treatment integrity:

Putting aside the studies with an “unclear” assessment, there are 89 studies (73%) where it has been possible to estimate the overall treatment integrity. Figure 2.3 summarises the overall pattern of treatment integrity across all 122 studies. 49 studies (55% of the studies where an estimate was possible) appeared to be above the 90% threshold, with a further 29 (33%) above the 60% threshold, of which 19 out of 29 (65%) were over 80%. There were 11 studies (12%) where the estimated treatment integrity fell below the 60% threshold and where there were clear indications in the reported evaluation that there was a high risk of bias.

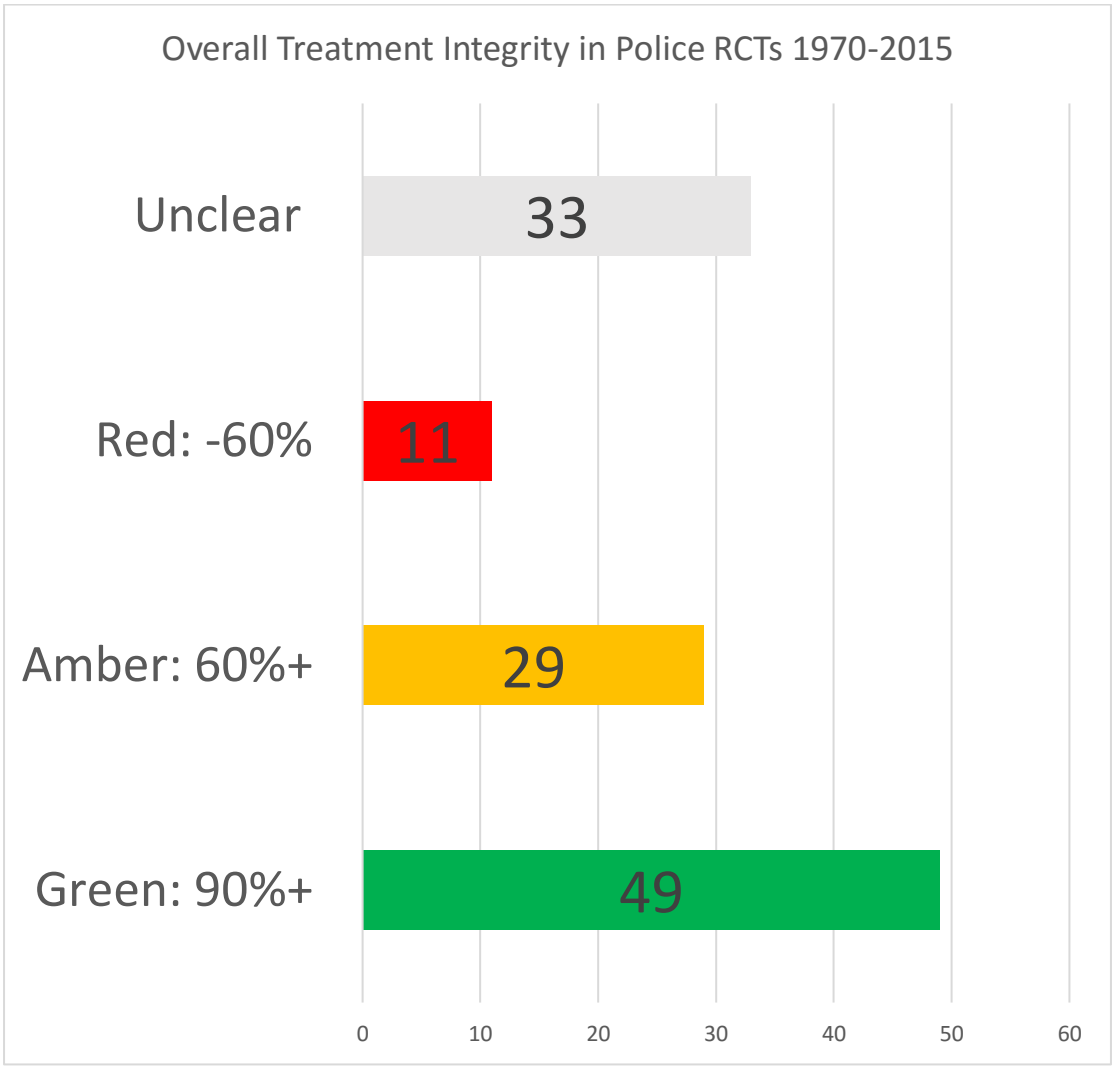


Figure 2.3: Overall Treatment Integrity in Police RCTs from 1970-2015 (n.122).

2.5.2(c) The "Red" estimates:

In Table 2.8 the "Red", below 60%, studies have been extracted from Table 2.4. Of the 11 studies, there are 5 in the area of juvenile justice, 2 of which (McCold and Wachtel, 1998a and b) have been categorized primarily as restorative justice trials on the basis that the primary hypothesis being tested was in the effectiveness of restorative justice conferences. However, all 5 shared some common design features. They were all trickle flow designs which were testing treatments with young offenders. There were, however, two distinct reasons for the low treatment integrity. The researchers in both the Bethlehem studies (McCold and Wachtel, 1998a and b) chose to apply the randomization process after selecting the population for the study but before the process of consent to participate in the study. The process of obtaining consent resulted in a substantial level of attrition from the sample in both the violence and property experiment. In the Memphis experiment within Dunford et al.'s (1982) National Evaluation of Diversion Projects, the researchers had a problem with the consistent presence of a key member of the field research team and had a high level of attrition from one part of the sample in the early stages of the trial.

In contrast, in the other two juvenile justice studies (Byles and Maurice, 1979 and Giblin, 2002) the major issue was the delivery of the treatment as intended to the treatment sample in the experiment. More than half of the juveniles and their families in Byles and Maurice's study either did not take up the treatment offered or disappeared from the trial. In Giblin's Anchorage CAN project the treatment required at least two visits per month from the police officers involved in the trial. Only 45% of the treatment sample received the required dosage and the author lamented that "the number of contacts received by each CAN Program juvenile varies and likely indicates that the program was only partially implemented for some probationers" (Giblin, 2002:128).

Indeed, with the exception of the McCold and Wachtel (1998a and b) and Dunford et al. (1982d), Table 2.10 provides a strong indication that, whilst there were some issues with the process of randomization and some problems with attrition from the sample, the more frequent and problematic risk was the failure to deliver the treatment as intended.

Earle (1973) provides a graphic illustration of the problem. In this very early trial, Earle, a senior executive in the Los Angeles Sheriff's Department, set up an experiment to test two models of training: a control sample who were treated to the "as is" military style "stress training"; a treatment group who were to be exposed to a "non-stress" model with less drill, greater staff-student interaction and respect and a more behavioural-science based class curriculum. Unfortunately, because the researchers had included former soldiers and police officers from other jurisdictions in the sample, the latter persuaded the treatment group that they were missing out on the "real" training and that they should voluntarily undertake all the stress elements that were missing from the treatment approach. Realising their mistake, the researchers changed the exclusion criteria for the second intake and tracked compliance with the treatment approach more carefully. The importance of achieving a high level of treatment integrity was evidenced by the results: with Intake 1 the performance of the treatment group was indistinguishable from the control; in Intake 2 there was a substantial and significant difference which the author interpreted as supporting the hypothesis under test.

Author(s)	Year	Coded RCT topic ((after Braga et al., 2014)	SAMPLE	UNIT	Treatment integrity %
Banerjee et al.	2013	Road Safety - Drink Drive	123	Roadside Breath Stations	53
Byles & Maurice	1979	Juvenile justice	305	Juveniles with 2 or more priors	45
Davis & Medina-Ariza	2001	Domestic violence	60	Housing Projects/Households	50

Dunford et al.	1982d	Juvenile justice	975	Juvenile offenders	54.35
Earle	1973	Police Training	174	Deputy Sheriffs	57.5
Giblin	2002	Juvenile justice	190	Juvenile offenders	45
Graziano et al.	2014	Citizen feedback/interaction	51	Beats in Chicago	47.5
Ireland et al.	2007	Health and Welfare	129	Police Officers	51.9
Lurigio & Rosenbaum	1992	Crime Victim outreach	122	Police Recruits	26.5
McCold & Wachtel	1998	Restorative Justice	111	juvenile offenders	31.5
McCold & Wachtel	1998	Restorative Justice	181	juvenile offenders	49.5

Table 2.8: Police RCTs with a below 60% estimate of treatment integrity (n.11).

2.5.2. (d) The "Green" estimates:

It is equally important to consider if there any conclusions that can be drawn about the experiments that appear to have achieved high levels of treatment integrity. This group (see Table 2.9) are spread across 18 different topics and across the whole period from 1970-2015. In amongst the "90%+" group there are a wide variety of study designs from trickle flow juvenile justice (2), restorative justice (8) and domestic violence (5) studies to place based hotspots studies (7) and large cohort DARE (4) and Legitimacy studies (4). With the last two categories, the DARE and Legitimacy studies, it is important to highlight the fact that whilst high levels of treatment integrity appear to have been achieved in the processes of random assignment and treatment delivery, these studies are more vulnerable to bias from post-test attrition as a result of non-completion of the surveys that are a critical part of their evaluation. For example, survey completion rates in Mazerolle et al. (2012) and MacQueen and Bradford (2014) were 13.16% and 6.6% respectively.

There is also a group of 8 studies, of which the Health and Welfare RCTs (5) represent the largest group, in which the study combines elements of both laboratory and field experiment. In these “field laboratory” studies, such as Wells et al.’s (2015) Eye Witness RCT and Sousa et al.’s (2010) Taser RCT, high levels of treatment integrity were achieved by testing the intervention with real police personnel but in a highly controlled environment. These RCTs present a considerable contrast to the complex real-world operational conditions of the juvenile justice (2), pre-court diversion (1 - Operation Turning Point), restorative justice (8), domestic violence (5) and hotspots studies (7), which, nevertheless, appear to have achieved high levels of treatment integrity.

Author(s)	Year	Coded RCT topic ((after Braga et al., 2014)	SAMPLE	UNIT	Treatment integrity %
Abrahamse et al.	1991	offender management	480	Offenders	91.25
Angel et al.	2014	Restorative Justice	192	Victims	99.47
Antrobus	2015	Crime Victim outreach	978	Crime Scenes	93.5
Clayton et al.	1996	DARE	31	Schools and Pupils in 6th Grade	93
Dunford et al.	1990	Domestic violence	330	Offenders in DV cases	92
Dunford	1990	Domestic violence	247	Offenders in DV cases	96.5
Esbenson et al.	2012	DARE	195	Classrooms	90
Gersons et al.	2000	Health and Welfare	42	Police Officers	97.6
Glick et al.	1986	shoplifting	1346	Shoplifters	92
Groff et al.	2005	Citizen feedback/interaction	314	Residents	98
Komro et al.	2004	DARE	24	Schools and 7th Grade pupils	98
Koper et al.	2013	Hotspots/crime places	117	Hot route sites	100

Ku & Blew	1977	Juvenile justice	36	Adolescent offenders	100
La Vigne & Lowry	2011	CCTV	50	Parking facilities	100
Langley	2014	Legitimacy	781	Passengers	98.4
Lu et al.	2012	Road Safety: traffic enforcement	80377	Private car owners	100
McCraty & Tomasino	1999	Health and Welfare	65	Police Officers	90.7
McEwen et al.	1986a	Differential Police Response	5510	Calls for Service	92.1
McEwen et al.	1986b	Differential Police Response	34795	Calls for Service	91.3
McEwen et al.	1986c	Differential Police Response	5497	Calls for Service	98
McGarrell & Kroovand	2007	Juvenile justice	782	Juvenile offenders	98.5
Mazerolle et al.	2012	Legitimacy	60	Random Breath Test Stations	100
Neyroud et al.	2015	Pre-court diversion	417	Offenders	91
Owens et al.	2015	Legitimacy	1444	Police Officers	100
Pate & Hamilton	1992	Domestic violence	907	Cases of Domestic Violence	90
Quinton	2011	Citizen feedback/interaction	7434	Citizens	100
Ringwalt et al.	1991	DARE	20	Schools and pupils	100
Roman et al.	2009	DNA/Crime detection	2150	Volume crime cases	99.9
Sahin	2014	Legitimacy	702	Drivers stoppped in speed checks	96.6
Santos and Santos	2014	Hotspots/crime places	48	Hotspots of crime	95.9
Shapland et al.	2006a	Restorative Justice	106	Offenders	92.2
Shapland et al.	2006b	Restorative Justice	186	Offenders	90.3
Shapland et al.	2006c	Restorative Justice	165	Offenders	95.1

Shapland et al.	2006d	Restorative Justice	165	Offenders	98.2
Shapland et al.	2006e	Restorative Justice	105	Offenders	95.7
Shapland et al.	2006f	Restorative Justice	105	Offenders	93.3
Sherman & Weisburd	1995	Hotspots/crime places	110	Crime Hot spots	91
Sherman & Rogan	1995	Hotspots/crime places	207	Blocks in Kansas City	97
Sherman et al.	1992	Domestic violence	1200	Domestic Violence cases	98.25
Shipley and Baranski	2002	Health and Welfare	54	Police Officers	100
Short et al.	1984	Health and Welfare	45	Police Officers	100
Strang et al.	1999a	Restorative Justice	900	Offenders	93
Slothower	2015	Crime Victim outreach	142	victims of crime	91
Sousa et al.	2010	TASER	64	Police Officers	100
Telep et al.	2014	Hotspots/crime places	42	Hotspots	100
Weisburd & Green	1995	Hotspots/crime places	56	Drug Hotspots	100
Weisburd et al.	2015	Hotspots/crime places	232	Police beats	100
Wells et al.	2015	Eye-witness identification	497	Line-ups	100
Wilson et al.	2001	Health and Welfare	62	Police Officers	95

Table 2.10: Police RCTs with 90% treatment integrity by subject category (n.49)

By combining the analysis of the “90%+” group with the analysis of the “Red” group with below 60% treatment integrity, it is also possible to draw out some tentative conclusions about the patterns of treatment integrity over time. The data in Table 2.10 has been sorted into decades, with the exception of the last half decade, 2010-2015, which, as we have seen, has been the most productive period of all. The Table shows that only 6 studies out of 17 completed before 1990 appear to have reached the 90% threshold. However, there is not a clear indication of increasing treatment integrity across the 45

years. The data suggests that the picture is a little more mixed. There appears to have been a reduction in the proportion of low integrity (below 60%) studies, at least in the most recent period. However, after a steady rise in the proportion of 90%+ studies, that same recent period appears to have seen a somewhat lower proportion (43% as against 52% for the period 2000-2009) of the high treatment integrity studies.

Years	No. of 90%+ RCTs	No. of RCTs below 60%	Total no. of RCTs	% of RCTs 90%+	% of RCTs below 60%
1970-1979	1	2	10	10	20
1980-1989	5	1	19	26	5
1990-1999	12	3	27	44	11
2000-2009	13	3	25	52	12
2010-2016	18	2	41	43	4

Table 2.10: Decade by decade numbers and percentage of Police RCTs with 90%+ or below 60% treatment integrity 1970-2015

There are a number of possible explanations for this. Firstly, the number of new scholars and institutions leading studies may have contributed. There are certainly a number of recent studies with authors who were not represented in Braga et al.'s sample (Owens et al., 2014 and Grossmith et al., 2015 for example) where the reporting of treatment integrity has been unclear and the studies have a narrative suggesting implementation problems. Secondly, the improved standards for reporting RCTs in journals such as the Journal of Experimental Criminology (where a CONSORT is normally required to support an article reporting an experiment) may mean that there has been more and better data available to this analysis compared to some earlier studies, where the methods and implementation sections are, at times, only a brief precursor to the reporting of the results.

2.6 Discussion of Findings

This chapter has explored two research questions: firstly, the size and nature of the universe of police RCTs prior to 2016; the level of treatment integrity in police RCTs. In response to the first, the search for police RCTs produced 122 Police RCTs completed and reported by 2016. The levels of treatment integrity have been set out in Table 2.4, but suggest that 78 of the 122 RCTs exceeded a 60% threshold, with 49 being above 90%.

Given the controversy over police RCTs that we have discussed above, it is an important finding that there have been a lot more Police RCTs than many scholars and practitioners have thought and significantly more completed by 2015 than Braga et al. (2014) found in their search completed in 2012. Yet, even so, compared to quasi-experiments and other quantitative studies with a control, they still represent a vanishingly small proportion of police research. The systematic search for the Global Police Database (Higginson et al., 2015) has identified more than 7000 studies, which would mean that RCTs account for less 2% of the quantitative studies in policing.

The core of the analysis presented above has attempted to provide estimates of the treatment integrity for the 122 RCTs that were discovered in the search. There are number of limitations to this process. Firstly, the process of estimation has had to rely on the information provided in a wide variety of formats across a half century in which the expectations of trial reporting have changed and, in many respects, improved. Secondly, whilst the Cochrane methods (Higgins and Altman, 2008) and the CONSORT format has provided a basis for the approach, none of the existing methods were designed to assess treatment integrity in the way intended in this study. Thirdly, the process has been applied to a body of RCTs which varies widely in design and topic. The relatively high level of “Red” rated RCTs in the field of juvenile justice appears to reflect both the study design – trickle flow – and the relatively challenging nature of the topic for study: issues to which we will be returning in more detail in Chapter 4.

Beyond the process of estimation, we have also acknowledged above that the process of searching for police RCTs presented some limitations. It is far from straightforward to find Police RCTs. Standard searching methods for meta-analysis are designed to search for studies of a specific intervention, rather than all the studies in a whole field as broad as policing. This highlights a further issue: police RCTs do not form a coherent, homogenous body of work. They cover an increasingly broad range of topics and have several different methods of controlled design. Hence, we have been cautious in making comparisons across the whole sample of RCTs.

Perhaps partly because of the challenges of assembling all the studies, there has, up to now, been no comprehensive repository of police RCTs that is readily available to scholars or practitioners, despite the important work of Lum et al. (2011) to develop a Matrix of crime prevention studies. Yet, despite this, RCTs in policing and criminal justice generate a level of controversy seemingly out of all proportion to their numbers (Sparrow, 2016). The principle reason for this lies in the central claim made for RCTs by their key advocates that they "allow researchers to assume that the only systematic difference between the control and treatment groups is the presence of the intervention; this permits a clear assessment of the cause and effects" (Mazerolle et al., 2014: 508). As such the randomized controlled trial would appear to be the preferred method – the so-called "gold standard" - when the purpose of the research and evaluation is to determine whether an intervention or treatment works.

The degree to which RCTs can be said to meet a gold standard is hotly contested. When Sherman (2009) advanced an argument in favour of "experimental criminology" supported by the widespread use of RCTs and systematic reviews, Carr (2009), Hope (2009), Tilley (2009) and Hough (2009) all contested the argument. However, their main arguments were not framed around contesting the internal validity of experiments. Hough, for example, accepted that "RCTs provide strong internal validity, but in complex settings offer weak external validity, making it hard to generalize from the experimental setting to other settings" (Hough, 2009:11).

The findings in this study suggest that there is a constant need to pay close attention to internal validity and, as Berk (2005) argued, start with an assumption that researchers reporting RCTs need to demonstrate more transparently that they have succeeded in sustaining high levels of treatment integrity in their study.

This study presents the first meta-analysis of the treatment integrity in police RCTs across the last 45 years. The findings suggest that whilst treatment integrity improved in the 1990's and 2000's, it may have fallen more recently at a time when more new scholars have been involved in conducting and managing police RCTs. Furthermore, it is clear that a much larger proportion of the research team and principal investigators have been "pracademics" – serving practitioners who are also academic researchers – or researchers operating within professional bodies rather than more traditional academic institutions or research foundations.

Balanced against this tapering of the improved levels of treatment integrity, there has been a reduction in the level of "unclear" studies from nearly half the early studies to a third of the more recent. This, taken together with the overall rise in treatment integrity from the early studies to present day and the significant increase in the number of the RCTs and the diversity of topics suggest that Clarke and Cornish (1972) were prematurely and overly pessimistic about the feasibility of experimental research in criminal justice.

The notable shift to insider, professionally led research is consistent with Shepherd's (2003) arguments for greater clinical, operational involvement in research. However, one possible conclusion from this analysis is that "novice" researchers operating within "novice" institutions may have found it more difficult to achieve high levels of treatment integrity. Feder et al. (2000), in reflecting on their experience, as external academics, of conducting their first RCT, commented:

> "Without enough thought, we assumed that running a randomized experiment would be similar to implementing any other research design. In this manner, we failed to take account of the rich literature in experimental research. By the time we discovered the literature on experiments, we were mired in many of the chronic problems that those before us had experienced and described in the experimental literature." (p.398).

As such, Feder et al. (2000) describe the "novice" experience in conducting and managing randomised trials in policing. Not only were they conducting their first RCT, the police agency involved in the study was new to experimental research and they were testing a new intervention. At just over 70% treatment integrity their RCT was, whilst by no means the most problematic of the 122 in Appendix 1, problematic enough from the research team to publish an article on the "lessons". Their analysis led them to the clear conclusion that researcher inexperience, the police agency's challenge in understanding the requirements of experimental research and the untested intervention all contributed to the problems that they had had in conducting and managing their study. It is to this, "novice theory", as an explanation of the level of treatment integrity in police RCTs that we now turn in Chapter 3.

3. "Novice theory" and treatment integrity in police RCTs

3.1 Introduction:

Building on the analysis in Chapter 2, this chapter explores a possible explanation or predictive model for the level of treatment integrity in policing RCTs: a three-pronged "novice" theory. Firstly, the theory states that a primary determinant of RCT integrity is the prior experience of the RCT's principal investigator(s) in running at least one RCT to completion, no matter how well conducted. The theory predicts that integrity levels will be higher where PIs are not novices than where they are. There are two further prongs to this theory: the experience of the agency in conducting RCTs and the extent to which the intervention has previously been field tested. The chapter explains and tests each of these prongs and concludes that the theory provides an important but not complete explanation for treatment integrity.

As "novices", Feder et al. (2000) were unusual, as researchers, in sharing the travails of their first RCT in policing. Another exception was Sherman, whose Appendix to his book on "Policing Domestic Violence" (1992) provides an earlier example of the challenges of new RCT researchers testing a new intervention in an untested police department. Sherman's analysis is particularly significant because of the contrast with his experience in the Milwaukee experiment, by which time he had become an experienced experimentalist. He subsequently distilled the experience of conducting 14 experiments in his chapter on "experimental criminology" (Sherman, 2010), in which he recommends strategies for overcoming the predictable "novice" risks: experienced, PhD qualified, principal investigators who should be actively directing the experiment on the ground; building a "field station" to conduct a series of experiments rather than initiating individual field experiments in greenfield agencies; rigorous pre-testing and piloting to ensure that new interventions are established effectively before formal evaluation.

However, most of the experiments in Appendix 1 were conducted before Sherman's 2010 chapter was published and several experiments had suffered

problems similar to Feder et al. (2000). Looking further back, it appears that 'novice theory" may be directly relevant to Clarke and Cornish (1972) with whose critique of RCTs we started Chapter 2. Clarke's curriculum vitae[2] shows that the Kingswood study was conducted when he was a relatively recently appointed Home Office researcher. The Kingswood study was also one of a very small number of RCTs that the Home Office conducted in the 1970s and one of the first completed and reported. There was only one UK-based police RCT in the same era, Rose and Hamilton's 1970 study for which the fieldwork was carried out in the North-West. Moreover, none of the Home Office publications listed in the back papers of Clarke and Cornish's Home Office report, and therefore published beforehand, appear to be a randomized experiment. Furthermore, the experiments that Clarke and Cornish reference in their bibliography are all North American. As such it appears that the Kingswood experiment can be viewed as both an experiment testing the effectiveness of the new treatment at Kingswood and an experiment in implementing a relatively new – to the UK criminal justice field at least – method of research. In this light, Clarke and Cornish's paper could be read just as appropriately as the lessons learnt in the field by scientists testing a new approach rather than the definitive rebuttal of the RCT method which it has been sometimes been interpreted as presenting (Smith, 2012).

It would seem important and timely to explore whether such a "novice theory" has a wider relevance to police RCTs, particularly given the expanding group of researchers that this study has discovered. Braga et al. (2014)'s found that a relatively small group of 126 individuals had been named as authors of 63 police RCTs. While they did not explore the relationship between treatment integrity and their author cohort, they did map the key relationships between the small core of experienced authors and the wider author sample. They found a clustering of researchers around a smaller group of ten relatively experienced RCT researchers. They did not, however, seek to breakdown

[2] http://rscj.newark.rutgers.edu/wordpress/wp-content/uploads/2016/03/Ronald-Clarke_March-2016-Vita.pdf

that clustering by topic, police agency studied or the prior experience of the research team in conducting RCTs in policing.

This chapter has been set out to provide an exploration of that task. It begins with the method adopted to populate the three columns in Table 3.1 below labelled “novice investigator team”, “new research station” and “novel topic or intervention”. The findings, which are summarised in Table 3.1, are then set out in more detail and analysed to test whether “novice theory” can provide a valid explanation for low and high levels of treatment integrity in police RCTs. Finally, the chapter concludes with a discussion of some of the alternative explanations. The research questions explored are:

(a) To what extent can “novice theory” explain low and high levels of treatment integrity in RCTs in policing?

(b) What are the implications of these findings for the conduct and management of police RCTs?

3.2 Exploring “Novice theory”:

The “Novice Theory” which we are exploring in this chapter can be simply stated as follows: that there is a higher risk of low treatment integrity in police RCTs in which new RCT researchers are testing a new topic or intervention in an untested police department or agency which has not previously been involved in the conduct or management of experimental research. The theory began to emerge as the author coded the 122 RCTs and examined the patterns of treatment integrity.

The most obvious method to explore this further would have been to survey all 122 research teams. However, the scale of this undertaking, combined with the search set out in Chapter 2 and the Case study in Chapters 4 and 5, would have been considerable. Moreover, as Dennis (1988) found in

surveying 40 author teams for detailed information about their RCTs, the logistical challenge would be immense. The improbability of achieving a credible sample was also emphasised by the experience of the author, during this research, in assisting a fellow researcher with an exploration of the learning from police RCTs, using the initial Matrix list from this research: the response rates were very low. From the author's own attempts to contact scholars for clarifying information, it seemed likely that the information returned would be heavily biased in favour of the experienced investigators in Braga et al.'s (2014) inner core rather than the novice group. Finally, with fifty years of police RCTs to analyse, it was quite apparent that the data would be biased to more recent RCTs in which the research teams were still active (and alive).

The approach adopted in this Chapter has, therefore, concentrated on desk research of the research team's CV's, their searchable publications, the bibliographies of their RCTs and drawing out the information from the articles and reports of the police RCTs in Appendix 1. This information has been converted into a simple binary judgment – Y for Yes and N for No - in the three right hand columns of Table 3.1.

3.2 (a) "Novice Researchers"

The "Novice Research team" column has been populated by reviewing the CV's and searchable publications of the listed authors of each of the police RCTs and the bibliographies of their RCTs. If this review demonstrated that any one of the authors had prior experience of a police RCT then the box has been coded "N". In addition, where the CV or publication research showed that an author had conducted a criminal justice RCT, such as a study of post-court diversion or youth justice, this was also coded "N". An example of this was Peter Greenwood, who was both a co-author in Abrahamse et al., (1991) and the principal investigator of the Paint Creek Youth Center RCT (Greenwood and Turner, 1983). In addition, where the author was supervised in their research by an experienced RCT investigator, this was also coded 'N'.

3.2. (b) “New Research Station”

Sherman (2010) developed the concept of a “field station” for police RCTs. He based the idea on the model deployed in agricultural research. He argued that the way forward for delivering high quality RCTs in the field in policing was to build a longer term social and experimental relationship between the researchers and an agency, in which a succession of experiments was conducted over time. However, until recently with the development of “field stations” in the West Midlands (which we will discuss more in Chapters 4 and 5), Western Australia and Queensland, most police RCTs have been delivered as a single study in an individual agency, for which we have used the term “research station” here to distinguish it from Sherman’s concept. Given that “novice theory” centres on experience and embedded knowledge of how to conduct and manage RCTs, the RCTs have been coded “N” against “New Research Station” only when there was evidence of a prior RCT being conducted in the agency within the previous five years.

3.2. (c) “Novel Topic or Intervention”

Coding for “novel topic or intervention” was the most straightforward of the three dimensions. However, there was still a need to make a judgment as to the replication of a topic, such as domestic violence or juvenile justice amounted to a new intervention or a replication which was sufficiently similar to a previously completed study. The replications of the Minneapolis Domestic Violence RCT have been treated as replications, because that was the clear intent of the studies, even though there were some variations in the design of some of the studies. The key issue in this case was the researchers had the opportunity to learn from the published lessons (Sherman and Berk, 1984 and Berk et al., 1988) of the original study. On the other hand, almost all the juvenile justice RCTs have been treated as “novel interventions” because each one of them had a bespoke treatment design.

3.2. (d) Limitations of the analysis

There are some limitations with this approach. For many of the older studies, the researchers have either died or retired and therefore their CV's were not readily searchable on the internet. Some of their publications may also not be searchable in an electronic form. Some researchers' CVs were very brief and did not provide a full list of all their published work. Even where their work was listed, it is possible that the titles of articles shown (as we discovered in searching for police RCTs) were not transparent in revealing RCTs. Furthermore, as we set out in discussing publication bias in Chapter 2, it seems more likely that failed experiments will have remained in the file drawer and the hypothesis being explored here would suggest that these experiments may have been more likely to have been managed by novice investigators. In summary, as with the estimates of treatment integrity set out in Chapter 2, the "novice" data should be treated as a best estimate from the information publicly available.

3.3 Findings: "Novice theory" in 122 Police RCTs:

The coding of the "novice" data is set out in Table 3.1. The Table provides four columns after each of the 122 RCTs: the treatment integrity estimate which has been colour coded to Green, Amber, Red or Unclear; Novice Investigator Y/N; New Research Station Y/N; Novel topic or intervention Y/N.

Author(s)	Treat as Assigned/treatment integrity %	Novice Investigator team	New Research station	Novel topic or intervention
Abrahamse et al. (1991)	91.25	N	Y	N
Ackerley (1986)	Unclear	Y	Y	Y
Amendola et al. (2011)	84.35	N	Y	Y
Angel (2014)	99.47	N	Y	N
Antrobus, E. and Pilotto, A. (2016)	93.5	N	N	Y

Ariel, et al. (2014)	Unclear	Y	Y	Y
Banerjee et al. (2013)	53	N	Y	Y
Banerjee et al. (2012)	84	Y	Y	Y
Bennett and Newman (2015)	Unclear	N	N	Y
Berk et al. (1992)	82	N	Y	N
Binder and Newkirk (1977)	Unclear	Y	Y	Y
Boyanowsky and Griffiths (1982)	Unclear	Y	Y	Y
Braga and Bond (2008)	Unclear	N	Y	N
Braga et al. (1999)	Unclear	N	N	Y
Byles and Maurice (1979)	45	Y	Y	Y
Clayton et al. (1996)	93	Y	Y	N
Davidson et al. (1977a)	Unclear	Y	Y	Y
Davidson et al. (1977b)	Unclear	N	N	Y
Davis and Taylor (1997)	84	Y	Y	Y
Davis and Medina-Ariza (2001)	50	N	Y	Y
Davis and Maxwell (2002)	Unclear	N	N	Y
Davis et al. (2007)	85	N	Y	N
Dunford et al. (1990)	92	N	Y	N
Dunford (1990)	96.5	N	Y	Y
Dunford et al. (1982)	84.7	Y	Y	Y
Dunford et al. (1982)	69.5	Y	Y	Y
Dunford et al. (1982)	78.2	Y	Y	Y

Dunford et al. (1982)	54.35	Y	Y	Y
Earle (1973)	57.5	Y	Y	Y
Eck and Wartell (1998)	83	Y	Y	Y
Esbenson et al. (2012)	90	N	Y	N
Gersons et al. (2000)	97.6	N	Y	Y
Giblin (2002)	45	Y	Y	Y
Glick et al. (1986)	92	Y	Y	Y
Graziano et al. (2014)	47.5	N	N	Y
Groff et al. (2005)	100	N	Y	Y
Groff et al. (2015)	Unclear	N	N	Y
Grossmith et al. (2015)	Unclear	N	Y	N
Hegarty et al. (2014)	88	Y	Y	N
Hirschel et al. (1990)	83.5	Y	Y	N
Ireland et al. (2007)	51.9	Y	Y	Y
Jennings et al. (2015)	Unclear	Y	Y	N
Jolin et al. (1998)	70.7	Y	Y	Y
Komro et al. (2004)	98	N	Y	N
Koper et al. (2013)	100	N	Y	N
Ku and Blew (1977)	100	Y	Y	Y
La Vigne and Lowry (2011)	100	Y	Y	Y
Langley (2014)	98.4	Y	N	Y
Lincoln et al. (1977)	Unclear	Y	Y	Y
Little et al. (2004)	64	Y	Y	Y
Lu et al. (2012)	100	Y	Y	Y
Lum et al. (2011)	84.6	N	Y	Y

Lurigio and Rosenbaum (1992)	26.5	N	Y	Y
MacQueen and Bradford (2014)	Unclear	Y	Y	N
Martin and Sherman (1986)	80	N	Y	Y
McCold and Wachtel (1998)	31.6	Y	Y	Y
McCold and Wachtel (1998)	48.6	Y	Y	Y
McCraty and Tomasino (1999)	90.7	Y	Y	Y
McEwen et al. (1986)	90.7	Y	Y	Y
McEwen et al. (1986)	91.3	N	Y	Y
McEwen et al. (1986)	98	N	Y	Y
McGarrell and Kroovand (2007)	98.5	Y	Y	N
Mazerolle et al. (2012)	100	N	Y	Y
Mazerolle et al. (1998)	70	Y	Y	Y
Mazerolle et al. (2000)	70	Y	Y	Y
Mejia et al. (2013)	Unclear	Y	Y	Y
Mohler et al. (2015)	Unclear	Y	Y	Y
Neyroud et al. (2015)	91	N	N	Y
Norvell and Belles (1993)	64.4	Y	Y	Y
Owens et al. (2014)	Unclear	Y	Y	N
Owens et al. (2015)	100	N	N	Y
Pate et al. (1985a)	Unclear	N	Y	Y
Pate et al. (1985b)	Unclear	N	Y	Y

Pate and Hamilton (1992)	89.9	N	Y	N
Pate et al. (1991)	61.3	N	Y	Y
Piza et al. (2015)	Unclear	Y	N	Y
Quay and Love (1977)	Unclear	Y	Y	Y
Quinton (2011)	100	Y	Y	Y
Ratcliffe et al. (2011)	Unclear	N	Y	N
Ridgeway et al. (2011)	68.2	N	Y	Y
Ringwalt et al. (1991)	100	Y	Y	Y
Roman et al. (2009)	99.9	Y	Y	Y
Rose and Hamilton (1970)	Unclear	Y	Y	Y
Rosenbaum et al. (1989)	80	N	Y	Y
Rosenbaum et al. (1994)	Unclear	N	Y	Y
Rosenbaum and Lawrence (2013)	Unclear	N	Y	Y
Rosenfeld et al. (2014)	Unclear	Y	Y	Y
Sahin (2014)	96.6	N	Y	N
Santos and Santos (2014)	95.9	Y	Y	Y
Shapland et al. (2006)	92.2	N	Y	N
Shapland et al. (2006)	90.3	N	N	N
Shapland et al. (2006)	95.1	N	Y	N
Shapland et al. (2006)	98.2	N	N	N
Shapland et al. (2006)	95.7	N	N	N
Shapland et al. (2006)	93.3	N	N	N
Sherman et al. (1989)	Unclear	N	N	Y

Sherman and Weisburd (1995)	91	N	Y	Y
Sherman and Berk (1984)	83	Y	Y	Y
Sherman and Rogan (1995)	97	N	Y	Y
Sherman et al. (1992)	98.25	N	Y	N
Short, et al. (1984)	100	Y	Y	Y
Strang et al. (1999)	93	N	Y	Y
Strang et al. (1999)	88.5	N	N	Y
Strang et al. (1999)	76.9	N	N	Y
Strang et al. (1999)	86	N	N	Y
Shipley and Baranski (2002)	100	Y	Y	Y
Skogan and Wycoff (1987)	85	Y	Y	Y
Sloboda, et al. (2009)	75	Y	Y	N
Slothower (2014a)	91	N	N	Y
Sousa et al. (2010)	100	N	Y	Y
Stratton (1975)	Unclear	Y	Y	Y
Tanigoshi et al. (2008)	85	Y	Y	Y
Taylor et al. (2001)	85	N	Y	Y
Taylor et al. (2011)	Unclear	N	Y	Y
Telep et al. (2014)	100	N	Y	Y
Weisburd and Green (1995)	100	N	Y	Y
Weisburd et al. (2015)	100	N	Y	Y
Weisburd et al. (2011)	78.57	N	Y	N

Weisburd et al. (2008)	Unclear	N	Y	Y
Wells et al. (2005)	84	Y	Y	Y
Wells et al. (2015)	100	Y	Y	Y
Wheller et al. (2013)	86.1	N	Y	Y
Wilson et al. (2001)	95	Y	Y	Y

Table 3.1: Treatment Integrity and "Novice theory" in police RCTs

Three sets of data are focused on in more detail below to explore the relationship between the three novice categories and treatment integrity: the relationship between the eleven "red" rated RCTs and the novice categories; the "green" rated RCTs and the novice categories; the treatment integrity scores for the RCTs where no "novice" or only one novice category were present.

3.3. (a) Red Rated RCTs and "novice" categories

Out of the eleven "red" rated RCTs, where the treatment integrity estimate was below 60%, 8 out of 11 or more than 70% were conducted by novice researchers, in a new research station and on a novel topic or intervention. Of the three exceptions, Graziano et al. (2014) was conducted in Chicago, which had recently been the research station for another study (Rosenbaum and Lawrence, 2013), by an experienced team but deploying a new intervention, which the researchers observed was only partially implemented. Likewise, Davis and Medina-Ariza (2001) were an experienced research team using a novel intervention that experienced implementation problems. Finally, Lurigio and Rosenbaum (1982) encountered implementation failures in Detroit, testing a novel victim oriented training package.

The relationship between the 'red' rated RCTs and the three novice categories would appear to be a strong one, with the strongest relationship centring on the challenges of implementing novel topics or interventions.

3.3. (b) Green Rated RCTs and “novice” categories

Of the Green rated RCTs with a treatment integrity estimate of over 90%, 13 out of 48 (31%) were “novice” in all 3 categories, 16 had 2 novice categories and 18 had one category, with only one RCT showing no novice categories.

The 13 Green rated studies, which have been coded “novice” across all three codes, but nevertheless achieved high treatment integrity, appear, at first sight, to contradict “novice theory”. However, there are some caveats to this appearance. Ringwalt et al. (1991), Quinton (2011), Roman et al. (2008) and La Vigne and Lowry (2011), although undertaking their first police RCTs, were members of research groups in the Research Triangle Institute, National Policing Improvement Agency and Urban Institute which had substantial experience in field research in policing or public services. In a similar way, the three health and welfare RCTs (Shipley and Baranski, 1999, Short et al., 1984 and Wilson et al., 2001) were conducted by students or academics at medical or psychological faculties with experience of conducting clinical experiments. Santos and Santos (2014) had access to support from the academic network researching place based policing (Santos, personal communication, 2016).

The six remaining studies fall into three groups: the first study was very small (Ku and Blew, 1977); two more were conducted in a highly controlled “field laboratory” environment (Lu et al., 2012 and Wells et al., 2015). The final group were the three RCTs in the multi-site study of Differential Police Response (McEwen et al., 1986). They were conducted by a research consultancy, the Research Management Institute (now Institute for Law and Justice). The RMI researchers, McEwen and Connors were the principal investigators. Although not experienced at RCTs, both appear to have had several years of prior police research experience before starting the project in 1980. In addition, Connors could be seen as part-“pracademic”, having been in the Department of Justice before becoming a researcher. They appear to have built effective partnerships in all three agencies engaged in the study,

which required them to tailor the intervention and randomisation approach to each agency's operating procedures.

In summary, although there are 13 green rated studies which appear to be "novice" studies across all three categories, 11 of these were conducted by research teams who had institutional or network experience of police research to draw on.

3.3 (c) RCTs with no "novice" or only one novice category

There were four RCTs with no coded novice category. These were part of the programme of restorative justice RCTs that Sherman et al. (2006) conducted in Northumbria and London. There were six, linked RCTs. The first 2 at each site have been coded "novice research station" and the remaining 4 as "no novice" to reflect the initial novelty and then subsequent continuity of experimentation. Using the lessons from the original programme of Restorative Justice RCTs in Canberra (Strang et al., 1999), the researchers followed Sherman's (2010) "field station" model of conducting a series of tests in the same departments. All the Northumbria and London RCTs reported a 90%+, green rated, level of treatment integrity.

In addition to the UK RJ RCTs, the RISE RJ studies (Strang et al., 1999) could be argued to be close to a "no novice" coding. Four RCTs were conducted in the same agency. However, as each was, at that stage, a novel intervention, they have been coded as such. All the RISE experiments had a relatively high, Amber, level of treatment integrity.

A further RCT, Weisburd et al.'s (2011) hotspot study in three police agencies came close to "no novice" coding, because one of the three sites – Redlands – had already been a trial site for Weisburd et al.'s (2008) study on risk focused policing. This is particularly significant because Famega et al. (2016) subsequently analysed the factors behind treatment integrity in the study and

reported a substantial site by site by difference in the level of treatment integrity.

There were, overall, 36 trials in which only one novice category was present. Taking the trials in which there was sufficient data to make an estimate of treatment integrity, there were 19 out the 25 trials (76%) in which there was a Green rating, with 6 Amber and one red rating (Graziano et al., 2014 – discussed above). In 21 out of the 35 the novice category was a new research station. In the remaining 15 it was a novel intervention or topic.

3.4. Discussion of the findings: "Novice theory" as an explanation for treatment integrity in police RCTs

The findings suggest that the three novice categories are correlated with more than two thirds of the "red rated" RCTs where the treatment integrity level fell below 60%. Equally, more than three quarters of the RCTs with either no novice coding or only one novice category had a green rating, with only one being rated red, because of problems with implementing the treatment. Furthermore, whilst there were 13 Green rated RCTs with all the novice categories coded, further analysis suggests that, although the researchers may have been novices in carrying out their first police RCTs, they either had access to support from an institution with experimental experience or had substantial prior field research experience in policing.

The importance of that network of institutional or expert support was highlighted by Braga et al. (2014) in their description of the expert network linked to the 63 RCTs that they researched. It is also apparent in both the Santos and Santos (2014) and McGarrell and Kroovand (2007) RCTs, in which novice research teams in new research stations (and testing a new intervention in the case of Santos and Santos) managed to achieve high levels of treatment integrity. Whilst the Santos' could draw on the network of place-based researchers, McGarrell and Kroovand were provided with

support and advice drawn from the RISE RJ studies (Strang, 2016: personal communication).

The examples of Santos and Santos and McGarrell and Kroovand highlight a significant limitation to the analysis approach that we have used. Without the personal communications from the Santos' and Strang, the important network support to both studies would not have been so evident. Whilst the findings from the coding appear to provide support for the importance of "novice theory" as an explanation for treatment integrity, the findings are indicative rather than conclusive.

Above all, the approach does not provide detail on the mechanisms underlying the novice categories. For this it is necessary to go back to "lessons learnt" studies such as Feder et al. (2000). Their commentary identifies how the problems that they encountered had a relationship with each of the novice categories: as novice RCT researchers they found, as they reached back into the literature, that they were encountering known and predictable problems with resistance to random assignment, field discretion and treatment fidelity; the senior managers, staff and partners in the new research station in which they conducted the experiment did not appear to understand or buy in to the needs of the field experiment; the novel treatment that they tried to implement was steadily diluted as the challenges and costs of implementing it became apparent to the agency. Sherman (1992) provides some similar themes from the experience of conducting the first domestic violence RCT (Sherman and Berk, 1984) in Minneapolis. Notably, the researchers encountered problems engaging key middle managers and found frontline discretion had accounted for relatively high levels of covert reassignment of cases between treatments. When Sherman conducted the replication in Milwaukee (Sherman et al., 1992), preventive approaches were put in place including, particularly, tight oversight and tracking by two key middle managers.

Sherman's application of experience in Milwaukee and, then, subsequently, in the RISE RJ (Strang et al., 1999) and UK RJ experiments (Sherman et al.,

2006), might suggest that, of all three categories, novice research teams would be more predictive of problems with treatment integrity than either new research stations or novel topics or interventions. To explore this further, it is helpful to examine a number of the topics in more detail:

- juvenile justice/pre-court diversion: there is a relatively high level of red or amber rated RCTs amongst the 17 juvenile justice studies.
- Restorative Justice: in contrast, there is a relatively high level of green rated high integrity studies across the 13 Restorative Justice RCTs
- Hotspot policing: the first hotspot policing study had a high level of treatment integrity, despite appearing to be a novel topic.
- Body Worn Video: the first four Body Worn Video RCTs had significant problems with treatment integrity.

3.4.1. Juvenile Justice

Of the 17 juvenile justice or pre-court diversion studies 50% were “unclear” because insufficient data was provided in the report and 6 out of the remaining 8 studies were amber or red (see Table 3.2 below). The novice coding suggests that most of the studies have had novice research teams, who have conducted studies with new or bespoke interventions in agencies with no prior experimental experience. Only one of the authors, Dunford, featured in Braga et al.’s list of the lead authors with one or more police RCT to their name. However, both of Dunford’s domestic violence RCTs were completed in the late 1980’s, sometime after the National Diversion programme, which was commenced in 1976. It appears from the authors’ CV’s that the National Diversion programme, an ambitious national programme with 11 studies, four of which were RCTs, was the first police RCT for principal investigators in the research team.

Author(s)	Treat as Assigned/treatment integrity %	Novice Investigator team	New Research station	Novel topic or intervention
Binder and Newkirk (1977)	Unclear	Y	Y	Y
Byles and Maurice (1979)	45	Y	Y	Y
Davidson et al. (1977a)	Unclear	Y	Y	Y
Davidson et al. (1977b)	Unclear	N	N	N
Dunford et al. (1982)	84.7	Y	Y	Y
Dunford et al. (1982)	69.5	Y	Y	Y
Dunford et al. (1982)	78.2	Y	Y	Y
Dunford et al. (1982)	54.35	Y	Y	Y
Giblin (2002)	45	Y	Y	Y
Ku and Blew (1977)	100	Y	Y	Y
Lincoln et al. (1977)	Unclear	Y	Y	Y
Little et al. (2004)	64	Y	Y	Y
McGarrell and Kroovand (2007)	98.5	Y	Y	N
Neyroud et al. (2015)	91	N	N	Y
Quay and Love (1977)	Unclear	Y	Y	Y
Rose and Hamilton (1970)	Unclear	Y	Y	Y
Stratton (1975)	Unclear	Y	Y	Y

Table 3.2: Juvenile Justice RCTs (n.17)

However, as we shall discuss in more detail in Chapter 4, juvenile justice RCTs also share a number of implementation challenges, which appear to have proven particularly difficult for novice researchers: the studies need to follow a trickle flow design which requires researchers to sustain a consistent, controlled approach to the research across an extended period; eligibility screening, random assignment and managing attrition from the samples have been problematic in almost all the studies; unless the experiment is a very small one with a very restricted eligibility (such as Ku and Blew, 1977), treatments have to be tailored to the individual offenders, which has presented problems attaining intended treatment dosages.

3.4.2. Restorative Justice

The Restorative Justice RCTs (see Table 3.3 below) have, by contrast with the juvenile justice RCTs, a much higher level of treatment integrity, with the standout exception of the Bethlehem RCTs (McCold and Wachtel, 1998), where the decision to obtain consent after randomization led to a high risk of attrition bias. They are also notable for being the only RJ RCTs in which there was a novice research team: McCold and Wachtel, neither of whom had previously led an experimental evaluation, did not follow the design being tested in the RISE studies (Strang, 2016: personal communication). In contrast, the eleven RISE and UK Restorative Justice Consortium studies, which represent all the other RJ studies listed, were led by Sherman, who had previously led domestic violence, hotspots and repeat offender RCTs, and Strang. Sherman et al. (2015) show how key lessons from Minneapolis and Milwaukee domestic violence RCTs were embedded into the RISE studies and how, in turn, the specific lessons on conducting RJ RCTs were then deployed in the UK studies.

Author(s)	Treat as Assigned/treatment integrity %	Novice Investigator team	New Research station	Novel topic or intervention
Angel (2014)	99.47	N	Y	N
McCold and Wachtel (1998)	31.6	Y	Y	Y
McCold and Wachtel (1998)	48.6	Y	Y	Y
Shapland et al. (2006)	92.2	N	Y	N
Shapland et al. (2006)	90.3	N	Y	N
Shapland et al. (2006)	95.1	N	Y	N
Shapland et al. (2006)	98.2	N	Y	N
Shapland et al. (2006)	95.7	N	Y	N
Shapland et al. (2006)	93.3	N	Y	N
Strang et al. (1999)	93	N	Y	Y
Strang et al. (1999)	88.5	N	Y	Y
Strang et al. (1999)	76.9	N	Y	Y
Strang et al. (1999)	86	N	Y	Y

Table 3.3: Restorative Justice RCTs (n.13)

However, it is not possible to draw a simple conclusion that the prior experience of the researchers in RISE contributed directly to the higher levels of treatment integrity in the UK. Sherman et al. (2015) outline how, in their view, the very different management and institutional structures played a key part in supporting a higher level of treatment integrity: in RISE the researchers were kept at arm's length and the police field management of the experiment

had high levels of turnover; in the UK, the researchers played a key role as facilitators of implementation and the police team was a small stable specialist group. This suggests that the extent to which a novice or experienced researcher can deliver high levels of treatment integrity is likely to be the result of the interaction between a series of factors, research skills and experience, the research topic and the institutional and organizational framework for the research programme.

3.4.3. Hotspots Policing

On the face of it, the high levels of treatment integrity in the first hotspot study (Sherman and Weisburd, 1995) could be seen as suggesting that novel topics do not necessarily present high risks. However, this would ignore the key patrol studies from the 1970's and, in particular, Kelling et al. (1974) and Boydstun, 1975), which were central to the design of the experiment as the article reporting the RCT makes clear. Sherman and Weisburd's analysis of the Kelling et al. (1974) experiment concluded that "a substantive bias towards the null hypothesis…may have been created by insufficient differences in patrol dosages" (p.627). Their analysis makes clear that this was, therefore, not a novel topic for experimentation and that they had designed their experiment to overcome the problems encountered in Kansas. Moreover, not only was Sherman, by this stage an experienced RCT researcher, but the research station for the study, Minneapolis, was a relatively small department with stable leadership that had already been engaged in experimental research (Sherman and Berk, 1984) and had been involved in developing the data necessary to understand hotspots in the run up to the experiment (Sherman and Weisburd, 1995). In Sherman's later model, Minneapolis was a developing into a "field station" (Sherman, 2010).

Moreover, a number of recent RCTs (Koper et al., 2013, Telep et al., 2014 and Weisburd et al., 2015) were carried out in previously untried departments and yet still showed high levels of treatment integrity on account of the detail devoted to the consistency of patrol dosage. In all three cases, experienced

researchers used new technologies – the GPS logs from police communications systems - to track patrol dosage. Wain and Ariel (2015) and Ariel, Weinborn and Sherman (2016) have demonstrated the importance of GPS technology as a means of tracking dosage fidelity. Equally, Ariel et al. (2012) have demonstrated how using a web-based randomizer could support researchers in controlling and tracking the randomization process. We will be exploring the extent to which technology can compensate for “novice theory”, particularly novel treatments, in Chapter 4.

The lessons from the hotspots studies are, therefore, like the RJ RCTs in that they are more complicated than a simple relationship between ‘novice theory” and treatment integrity. Experienced researchers, particularly with the support of new methods to ensure effective tracking, secured high levels of treatment integrity in new research stations. One additional factor in several these may well be the involvement of practitioners as part of the research team. Sherman (1992) highlighted the significance of the practitioner contribution to the high levels of treatment integrity in Milwaukee. In the Sacramento study (Telep, Mitchell and Weisburd, 2014) was driven from within the department by Mitchell, a Sergeant and PhD student, whilst Santos and Santos (2014) combined researcher and senior manager. Rachel Santos commented “in both of our experiments, Roberto (my husband) who is a commander in the PD, has a PhD, and worked directly for the chief, oversaw both experiments. While the second, more recent one involved the entire agency, they had a systematic process already in place, so it was a matter of "pulling" the patterns from response randomly… I know it would have been difficult for me as an outside researcher, especially because I wouldn't have been their supervisor…” (Santos, R., personal communication, 2016).

3.4.4. Body Worn Video

Novel treatment is the dominant theme of the most recent group of RCTs – four Body Worn Video (BWV) RCTs – which were all carried out by new experimenters in policing. By the very nature of the relative novelty of the

technology, they were novel experiments and three out of the four were carried out in police forces that had not previously engaged in RCT experimentation. None of the four experiments provide clear enough evidence in their reports to be able to assess their treatment integrity with any confidence. Ariel et al. (2014) concluded that they did "not know how well the requirements were implemented, and it is difficult to estimate the fidelity of the intervention" (p.22). Ironically, although the purpose of the trial was to test the effectiveness of a visible form of accountability on officers' discretion to use force and citizen interactions with officers, the researchers found accounting for the use of the mechanism to be problematic.

The Rialto experiment has since become the "trail-blazer" for a series of BWV RCTs (Lum et al., 2015). It also had some other unique characteristics. One of the principal investigators, Farrar, was the Police Chief of Rialto, a medium sized California police department. Although three other previous experiments (Earle, 1973, Sherman et al., 1992 and Braga et al., 1999) had involved senior police closely in the research process, only in Rialto was the Police Chief a principal investigator. As the authors commented "Rialto is a small force with a dedicated Chief who has directly managed the experiment" (p.22). They suggested that there was a need to test BWV in police forces where the experiment was being managed by a middle manager to see if there was a "Farrar" effect in Rialto.

Drover and Ariel (2015) documented just such an experiment in Wolverhampton in the West Midlands Police area in the UK. Drover, a police Chief Inspector, led the implementation of the experiment, with support from Ariel, from the middle of a large UK metropolitan force. In some respects, Drover's span of authority was greater than Farrar's. He had more officers and a larger area, but as a middle manager he also had the challenge of managing upwards as well as down. However, unlike Rialto, West Midlands was, when this BWV experiment was initiated, an experienced "field station" with three recently completed RCTs and strong senior management support behind a strategy of evidenced-based policing. Even so, Drover and Ariel (2015) report that by the end of the test period "78% of officers were using the

cameras for more than 10% of the time that they wore them", despite a well-documented and intensive tracking and supervision regime.

The BWV trials in Essex, London and Orlando each appear to suggest that an experimental design that is heavily reliant on compliance by individual officers in the high discretion environment of field operations may be particularly difficult to deliver with high levels of treatment integrity. As we have seen above, none of these studies provides convincing data on attrition or treatment levels. In contrast, Drover and Ariel drew on lessons from and, in Ariel's case, experience gained, in a number of the most recent hotspots studies in which high levels of compliance with patrol dosage in experimental areas that had been achieved by detailed tracking, supported by technology and individual feedback through supervisors (Telep et al., 2014, Ariel and Smallwood, 2015 and Ariel, Weinborn and Sherman, 2016). They developed a model of implementation which drew on both the lessons of RCT implementation and the wider leadership, change and implementation literature, particularly Kotter (2012) and Fixsen et al. (2005).

3.4.5. Summary:

In each of these four topics – Juvenile justice, RJ, Hotspots and BWV – it appears that "novice theory" whilst providing a useful explanation for the overall patterns of treatment integrity does not provide a sufficient explanation for specific RCTs. There was a high level of novice coding across the juvenile justice RCTs but it is also a topic area that presents some significant challenges in design and implementation. On the other hand, the RJ RCTs, which had very similar challenges seem to demonstrate what can be achieved with a combination of experienced researchers and using the lessons from prior RCTs. The Hotspots RCTs indicate that new technologies and practitioner involvement in the research team may provide further support for high levels of treatment integrity. Yet, the first four BWV RCTs show that implementing technology innovation in policing can also present challenges.

3.5 Further explanations for treatment integrity in police RCTs?

Drover and Ariel's approach to the challenge of RCT implementation raises the further question of how far the problems with achieving high levels of treatment integrity are specific to the delivery of police field experiments and how far they are related to more general lessons about the implementation of change and innovation within and outside policing. There is a very substantial literature that has documented the problems of implementing change and innovation in policing. Weatheritt (1986) analysed a set of local innovations in UK policing in the early 1980s and found that senior leadership change rhetoric was not matched by either detailed attention to change management or frontline buy in to the approach. Lum et al. (2012) argued that frontline officers failed to engage in the process of innovation because they saw such changes as fads which were quickly discarded as senior leaders moved on. Skogan (2008) identified the impact of resistance from middle managers, which is also evidenced in Sherman's analysis of the Minneapolis DV RCT in which strategies to bypass the middle and deal directly with the frontline were important (Sherman, 1992).

Both leadership and middle managers were important in Chan's seminal analysis of changing police culture, but she also emphasized the importance of and interaction between the field – the external and internal structural conditions of policing - and the habitus or content of cultural knowledge. Whilst Chan's focus was on how far the police could change their relationship with ethnic minorities through training and recruitment strategies, her attention to the importance of changing assumptions about knowledge and its relationship with practice seems to be highly relevant to field of police experiments.

Police RCTs are not just change and research programmes, but also a process of change that directly challenges existing practice and knowledge by testing an alternative hypothesis in the field. All the key experiments over the last 50 years have, initially at least, tested an innovation that contradicted accepted knowledge and practice and has done so with a method,

randomisation, that Greene (2013) has argued, is antithetical to deeply held assumptions about equity of treatment in criminal justice. As such, Chan's work not only emphasizes the difficulty of the change process in police RCTs, but also the likelihood that successful conduct and management of the trial necessarily requires careful attention to both dimensions: the field, including the governance, strategic leadership, middle management, tracking and supervision, performance management and evaluation process; the habitus, including attention not just to training the process of the experiment, but also to developing the understanding of the wider framework of evidence-based policing and experimentation. Drover and Ariel's (2015) model for implementing a BWV RCT paid attention to both dimensions.

Moving beyond policing, Fixsen et al. (2005), Greenhalgh et al. (2005), Stith et al. (2006) and Durlak and Dupre (2008) have conducted meta-analyses of the factors affecting implementation of programmes in a wide range of disciplines. All the studies were focused on understanding implementation in "real world settings by *non*-researchers" (Durlak and Dupre, 208:328), whereas police RCTs tend to rely on the implementation of innovations conducted and managed by researchers and developed in conjunction with a field experiment. However, as we have seen, more practitioners are now engaged in the research process alongside their more traditional role in implementation, which might suggest some erosion of the differences between the two models. The most recent of the studies, Durlak and Dupre (2008) provided a composite framework of factors that influence effective implementation which would appear to be highly relevant to both models:

- Innovation characteristics: the findings we have set out above suggest that different RCT topics with different "innovation characteristics" would appear to present some distinct challenges to sustaining high levels of treatment integrity. Durlak and Dupre highlight two dimensions: the importance of compatibility with the host organisation's mission, priorities and values; adaptability or the extent to which the innovation can be tailored to the local context, culture and needs. We have set out a number of examples above in which the researchers

reported the importance of senior management support and, therefore, a coincidence of experimental and organizational priorities. However, adaptability is more challenging for experimentation, particularly where the study is a replication. Dunford (2000) suggested that the process was more akin to negotiation, which would accord with Strang's (2012) description of the constant maintenance required to sustain the "coalition" of researchers and practitioners through the experiment.

- Provider characteristics: The "provider" would seem capable of being interpreted as the researchers in the case of RCTs. Durlak and Dupre identified a number of provider characteristics of which "skill proficiency" was the most strongly identified across the other key meta-analytic studies that they examined (Fixsen et al., 2005, Greenhalgh et al., 2005 and Stith et al., 2006). The findings in this research suggest that there may be a relationship between high levels of treatment integrity and the experience or novice status of the research team in conducting and managing police experiments. In a similar vein, Durlak and Dupre (2008) concluded that providers who "feel more confident in their ability to do what is expected (self-efficacy), and have the requisite skills are more likely to implement a program at higher levels of dosage or fidelity" (p.336).
- Community factors: aside from funding, politics and policy considerations, which are all relevant to police RCTs, Durlak and Dupre stressed the importance of the prevailing body of knowledge in the field. In their model that knowledge provided the supporting and reinforcing context for the innovation being implemented. Although we have not identified this issue in the factors related to treatment integrity, Sherman (2010) emphasised the importance of a credible hypothesis as a key element of the research motivation which he argued was a key factor in both securing the initiation and sustaining an experiment.
- The prevention delivery system or features relating to organizational capacity: Durlak and Dupre divided these factors into general organizational factors, specific practices and processes and specific staffing considerations, including leadership and internal champions.

We have identified that there may be an association between high levels of treatment integrity and police departments that have experience of experiments. Durlak and Dupre pointed to the importance of the "norms held by an organization in relation to its willingness to try new approaches" (2008: 337). Sherman (2010) used similar language in describing the characteristics of an effective "field station", which he argued was a key factor in successful experiments. Sherman (1996 and 2010) also emphasized the critical importance of leadership and championing of the experiment.

- The prevention support system: Durlak and Dupre focused on the importance of training and technical assistance, which are both functions that benefit from experience and technical expertise. They identified the importance of an approach to both which was at once active and engaged and continuous. The paucity of detail in the accounts of the training and support provided in many of the police RCT reports would prevent any credible comparative analysis of this set of factors. However, Sherman (2010), Slothower, Sherman and Neyroud (2015), Sherman et al. (2015) and Drover and Ariel (2015) all provide strong endorsement of the importance Durlak and Dupre's active model of training and expert feedback.

There seem, therefore, to be some strong common themes as between meta-analyses of programme implementation and the analysis of police RCTs presented here. Three themes, which appear to have an important relationship with both treatment integrity in police RCTs and implementation fidelity in innovation programmes, stand out: the importance of "provider" or researcher experience; the innovation culture of the organization or "research station"; the innovation characteristics of the programme. However, as Chan's (1997) study of change management in policing reinforced, these themes need to be understood within the field and habitus of policing when considering police RCTs. The relationship between police and researchers may be changing. There is certainly some indication of this in the growing involvement of police officers in the process of experimental research, but even so, as Alpert et al. (2014) have shown, few police-research partnerships

have deep roots. Until the last decade the research motivation (and funding) for the majority of the police RCTs has been external to the organization and often driven by a reform agenda (Braga et al., 2014).

3.6 From Novice to Expert?

There is a well-developed literature in fields of medicine and education on the stages of skills acquisition, which offers some important insights on “novice theory”. Dreyfus and Dreyfus (1980) set out a five-stage model of progressive proficiency: novice; competence; proficiency; expertise; mastery. This model was applied and adapted by Benner (1982) to the acquisition and development of skills in nursing. Benner suggested that the process could be characterised as progressing from novice, where practitioners were taught rules to guide action, to expert in which knowledge and field experience combine to develop a “deep understanding” of both the context and the appropriate courses of action (p.406). Feiman-Nemser and Remillard (1996) applied the stage model to teacher development and summarised it as “an initial stage of survival and discovery, a second stage of experimentation and consolidation and a third stage of mastery and stabilization” (p.66).

There has been considerable debate of Dreyfus and Dreyfus (1980)’s original staged model and Bonner’s development of it. Dall’Alba and Sandberg (2006) summarise the main concerns with such staged models as their tendency to direct attention away from the nature of the skill being developed and the importance of “understanding of, and in, practice” (p.388). Dall’Alba and Sandberg (2006) also drew attention to the fact that not all novices developed by stages into experts. In turn, this highlights the potential significance, identified by Braga et al. (2014), of a network of novices supported by a smaller group of super-experts able to guide and develop the practice of those learning.

As we have seen above, in applying the staged model to novices in field experimentation, there appear to be two key issues worthy of comment. Firstly, it is necessary to consider not only the novice status of the field

researcher, but also the context – the novice agency and the novel intervention. Secondly, as Feder et al. (2000) discovered, being an expert in wider forms of research did not necessarily translate into being an expert in field RCTs. Indeed, Feder et al. (2000) found themselves repeating errors that previous RCT researchers had identified.

3.7 Conclusions: "novice theory": a partial explanation of treatment integrity?

The analysis and discussion set out in this chapter has explored two questions: the extent to which "novice theory" provides an explanation for the high and low treatment integrity in police RCTs; the implications of the findings for the conduct and management of Police RCTs.

On the first question, it appears that "novice theory" can provide an explanation for the general patterns of treatment integrity. However, a more detailed examination of four topics – Juvenile Justice, RJ, Hotpots and BWV – suggest that there are other factors which may also be important in determining the effective conduct and management of police RCTs and, thereby, the treatment integrity of the study.

This suggests that the second question – the implications for the conduct and management of police RCTs – requires more detailed exploration of the process of experimentation. Feder et al. (2000) reflected that there were known and documented lessons from previous RCTs that might have helped them – novice researchers in a new research station seeking to implement and test a novel treatment – to overcome their novice status. The analysis of police RCTs in Chapters 2 and 3 has identified that there is now a growing involvement of practitioners in the research process. Many of these may be engaged in the research as part of a post-graduate qualification and, therefore, have supervisor or co-investigator support to balance their novice status. Even so, despite their practitioner, insider knowledge, they are likely to be novice field researchers. From the list of "in-flight" RCTs, there is also an

expanding group of researchers, many of them novice researchers on police RCTs, undertaking police RCTs. This broadening of the evidence-based movement suggests that it is important to broaden and modify “novice theory” with the lessons Feder et al. (2000) started to identify. In the next two chapters, we will seek to do just that by an in-depth analysis of a case study, Operation Turning Point.

4. Understanding treatment integrity in police RCTs: a case study of Operation Turning Point.

4.1 Introduction

The examination of treatment integrity in 122 police RCTs (Appendix 1) set out in Chapter 2 has provided some important emerging conclusions about the challenges of implementing trials effectively in policing to test a range of interventions. The analysis has shown a wide range of treatment integrity across 25 different topic areas. Chapter 3 has showed that what we have labelled “novice theory” – research teams new to police RCTs, new agencies hosting the research and novel topics or interventions – can provide an explanation for the patterns of low treatment integrity. However, the analysis in Chapter 3 has also demonstrated that there are other factors that contribute, including the research design, particularly the trickle flow designs required in studies such as pre-court diversion studies.

Yet, even here, the evidence is not consistent: the experiments testing some topic areas, such as restorative justice, have reported consistently higher levels of treatment integrity. In contrast, some topics such as Body Worn Video and Juvenile Justice appear to have consistently low levels. This chapter will seek to develop the analysis of treatment integrity in Police RCTs further by developing a model of the factors that appear to contribute to high and low integrity and using a case study, Operation Turning Point to test the model in more depth. The approach taken is more qualitative than Chapters 2 and 3.

In the first part, the chapter focuses on the topic areas which showed consistently low integrity in Chapter 2. Three topic areas, Hotspots, Body Worn Video and juvenile justice, have been reported with either incomplete data or low levels of integrity. In the hotspots studies, we identified that the major problem contributing to an “unclear” assessment was the lack of detailed reporting of the patrol dosage. The trials testing Body Worn Video are recent, the technology relatively novel in the field and it appears that the

operational challenges of tracking implementation of the treatment are the principal reason for the reported issues with treatment integrity (Ariel et al., 2014 and Drover and Ariel, 2015). Juvenile justice experiments, on the other hand, span more than four decades and have tended to report a range of significant problems with both sample attrition and treatment delivery.

Over those four decades, there has been increasing recognition of the importance of setting standards for the completion and reporting of trials. Martinson (1974) had highlighted problems with the methods and standards of reporting of some of the studies that he had reviewed. More recently, the Society for Prevention Research identified that there were different standards being used to identify and list programs and policies that have been tested and shown to be efficacious or effective (Flay et al., 2005). They commissioned work to develop an appropriate set of standards. Initially, Flay et al. (2005) and then, in an updated version, Gottfredson et al. (2015), argued, in describing those standards, that attention to the detail of attrition and treatments are critical in enabling accurate assessments of the efficacy and effectiveness of "evidence-based interventions" (EBI). Furthermore, in considering the potential for "scaling up" an intervention, whilst the random assignment process may not be so relevant, the "real world" issues encountered in delivering treatments are likely to be important (Spoth et al., 2013). Developing our understanding of these issues in police RCTs, therefore, affects the potential for EBI's to be disseminated across the field. Given the accelerating growth of an evidence-based policing movement, which we have documented above (Chapter 1), the justifications for pursuing these issues are increasing in importance.

There are a number of approaches that could be taken to developing the analysis in Chapters 2 and 3 further. Dennis (1988) selected 40 RCTs from his wider search and approached the investigators with a detailed questionnaire. Dennis asked the principal researchers for details of the issues and problems with random assignment that they had encountered. From this data he constructed a framework of six major risks to effective treatment integrity in criminal justice RCTs (Dennis, 1990). However, this approach, as

Dennis found, would tend to restrict the approach to recent RCTs in order to facilitate contact with the investigators. We identified in Chapter 2 that there have been some important changes over time that would be difficult to track with such an approach.

A second approach to exploring beyond the high-level analysis in Chapter 2 would be to start with a more detailed examination of the sample of 17 juvenile justice RCTs (Table 3.2) in order to identify common factors that appear to correlate with low integrity. However, a reading of the articles and reports from these trials would only take the analysis so far. A significant proportion of the studies are from the 1970s. The authors were primarily concerned to report the findings rather than the details of their method and implementation. The type of reporting that Flay et al. (2005), Perry et al. (2010) and Gottfredson et al. (2015) recommend, which would enable a more detailed comparison of risks of attrition and performance bias, is generally not provided. Nevertheless, there is sufficient detail in most to draw out some common themes with the processes of eligibility screening, random assignment, treatment and tracking methods. As such, the analysis of juvenile justice RCTs appears to provide an appropriate starting point.

For greater detail, there are a number of authors who have led or evaluated police RCTs who have published case studies or provided more detailed commentaries on the implementation of their studies. Earle (1973), Berk et al. (1988), Sherman (1992), Dunford (1990), Feder et al. (2000), Sherman et al. (2014), Sherman et al. (2015), Famega et al. (2016) and MacQueen and Bradford (2016) have all documented the problems that they have encountered and the approaches that they took to resolving them. These case studies and commentaries cover a range of police RCT topics and RCTs with both higher and lower integrity. Reviewing and analyzing them provides a second opportunity to explore factors that may correlate with both high and low integrity.

A third and potentially complementary approach is to follow the example of this last group of authors and document and analyse a case study. Given that

the analysis above has identified issues with treatment integrity in police RCTs testing juvenile justice a case study describing and analysing a pre-court diversion experiment would seem to be particularly relevant. The case study which will be set out below describes such a RCT focused on pre-court diversion of both juveniles and adults – Operation Turning Point – which has produced a relatively high level of treatment integrity (over 90%). Moreover, like Earle's (1973) trial, Turning Point also provides an example of a trial that had to be restarted after treatment integrity problems were identified in Stage 3, the first intended evaluation phase (Figure 4.1).

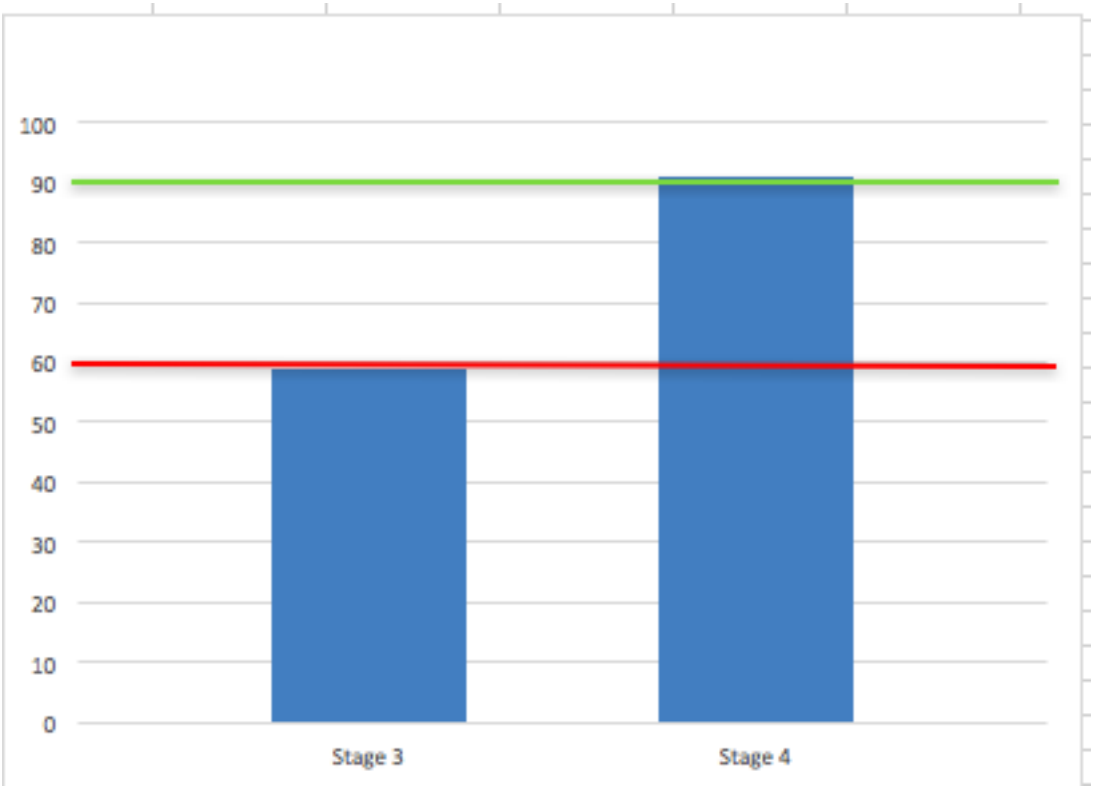


Figure 4.1: Treatment Integrity levels in Operation Turning Point Stages 3 and 4 (with 60% and 90% thresholds marked).

This chapter will, therefore, be structured around two main sections: a review of the issues in treatment integrity in the published case studies and commentaries and in police RCTs centred on juvenile justice; a detailed case study of Operation Turning framed against the issues drawn from that earlier analysis. In the final part, Part 3, the discussion of the analysis in Parts 1 and 2 will seek to draw out the lessons and focus on the following research questions:

What lessons for achieving "high integrity" can be derived from an analysis of published case studies of police RCTs and analysis of completed juvenile justice RCTs?

Using the framework from that analysis, what can a case study analysis of Operation Turning Point, a "high treatment integrity" experiment, tell us about the factors that might contribute to the conduct and management of successful (above 90% treatment integrity) randomised controlled trials in policing?

How might those factors add to or modify the "novice theory" proposed and analysed in Chapter 3?

4.2 Police RCTs: managing random assignment and treatment delivery:

We have argued above that the analysis of treatment integrity of police RCTs in Chapter 2 can be divided into two connected parts: random assignment attrition and the risks of bias resulting from cases being reassigned between treatment conditions or lost to the experiment; treatment performance which reflects the extent to which treatments intended were actually delivered at the duration and dosage specified. We will follow the same approach in reviewing the case studies and commentaries and in the next section reviewing juvenile justice RCTs. Moreover, the case studies and commentaries tend to fall into two groups: Earle (1973), Berk et al. (1988), Dunford (1990), Sherman (1992) and Sherman et al. (2015) are more focused on the process of random assignment, including eligibility screening and overt and covert overrides; Feder et al. (2000), Sherman et al. (2014), MacQueen and Bradford (2016) and Famega et al. (2016) are more concerned with the issues encountered in delivering treatments.

4.2 (a) Random assignment, eligibility screening and overrides:

The case studies suggest that it is important, in considering the potential risk of 'attrition' to start with the decisions and processes that precede random assignment. Earle (1973) and Dunford (1990) both identified the importance of pre-testing the randomization process and overcoming ethical, legal and practical challenges to the eligibility of cases entering the experiment. In Earle's trial the main problems stemmed from the inclusion of "contaminating" experienced ex-officers and former military personnel into a sample of otherwise new, civilian recruits. Their subsequent behaviours corrupted the treatment and led to the restarting of the trial. On the other hand, Dunford (1990), drawing on the issues that arose in the National Diversion Evaluation (Dunford et al., 1982), highlighted the importance of unearthing legal and ethical objections to the random assignment process from the outset and designing in "overt" reassignment processes to allow discretionary overrides where officers were unwilling to assign particular offenders to diversion. Test and Burke (1985) had already demonstrated, in a mental health triage experiment, that "practitioners often resent and consequently reject or sabotage random assignment because it bypasses clinical judgment" (p.91). Dennis (1988) suggested that this tendency required researchers to put in place overt reassignment protocols, which allowed officers to exercise discretion over eligibility within pre-agreed limits. There should not only be put in place, but practitioners had to be engaged in their design in order that they had ownership in and a clear understanding of the rationale for the experimental approach.

4.2. (b) "trickle flow designs"

Berk et al. (1988), Sherman (1992) and Sherman et al. (2015) discussed the challenges of the random assignment process in a series of domestic violence and restorative justice experiments. All of these were "trickle-flow" or "trickle process" random assignment designs in which the sample was assigned to treatments one at a time over an extended period (Riecken et al., 1974 and Shadish et al., 2002). All the juvenile justice and pre-court RCTs were also "trickle flow" designs. Sherman et al. (2015) characterized such trials as

reliant on the “hydraulic pressure” of cases from frontline staff – a pressure that required constant maintenance to sustain. Strang (2012), drawing on the experience of the Restorative Justice experiments, reflected on the complexity of maintaining the “coalition” of practitioners and agencies in order to keep the “hydraulic pressure” high through the period of the trial.

Quite apart from the challenge of maintaining case-flow, Dunford (1990) and Braucht and Reichardt (1993) demonstrated that “trickle flow” experiments were also more vulnerable to problems in the random assignment process than other designs. Braucht and Reichardt observed that “the corruption of the random assignment procedure appears to occur relatively often in field studies when random assignment is implemented via “trickle”” (p.80). Braucht and Reichardt proposed a typology of strategies for implementing random assignment in such designs:

- **the open-list method**: the random ordering is generated by the researcher and the sequence listed with blank entries for the staff to allocate to cases. This process – a version of which was deployed in the Minneapolis Domestic Violence experiment (Sherman and Berk, 1984) - requires very tight tracking and oversight to ensure that staff do not allocate cases to the treatment that they feel the case merits by simply skipping a line (Sherman, 1992).
- **the sealed-envelope method**: the treatment conditions are generated and recorded in sealed envelopes. Test and Burke (1985) demonstrated that this approach had similar problems to the first. In their trial, some staff opened the envelopes in order to select treatments that they felt were appropriate.
- **the research office method**: To overcome problems with relying on operational staff to carry out the assignment using researcher generated sequences, the research office method requires the staff to ring through to a researcher who allocates the treatment from random sequences. This approach was adopted in the Milwaukee Domestic Violence experiment (Sherman et al., 1992). It still requires careful

tracking, but with a matching focus on the pre-random assignment, street triage by which cases are defined as eligible (Sherman, 1992). Sherman (1992) contrasted the experience between Minneapolis and Milwaukee and concluded that it was important, in trickle-flow experiments, to separate the random assignment, as far as possible, from operational discretion. He subsequently recommended a protocol that "separates random assignment from operating staff" (Sherman, 2010:419). In particular, such separation was the only certain means to ensure the allocation sequence was blinded to the treatment providers, who could and demonstrably had manipulated allocation sequences in the past (Shulz et al., 1995).

- **computerized method**: Braucht and Reinhardt (1993) envisaged the computerized model as a means of enabling the research office to be brought to the field. They argued that the approach was capable of overcoming the problems with the alternative methods, ensuring blinded allocation sequences and providing a secure system that only a determined hacker could corrupt. Such a system had already been deployed in the three Differential Police Response RCTs, in which the random assignment was programmed in to the Computer Aided Dispatch system and treatment instructions could only be viewed once the cases had been registered in the system (McEwen et al., 1986).

Despite the relatively low levels of attrition reported by McEwen et al. (1986), Sherman remained convinced that separation was still the most effective way to ensure that random assignments were "above suspicion" (2010:419). The consequence of this was ongoing costs of sustaining a field staff over an extended time period (and often 'out of hours' as well) to manage allocations in trickle-flow studies. However, with web technology developments, Ariel, Sherman and Vila (2012) suggested that a web-based portal, which combined eligibility screening and random assignment with case by case tracking, could offer a means of field randomization which would allow the "treatment providers to do the random assignment process themselves, in a secure and friendly process" (p.195).

4.2. (c) The Cambridge Randomiser

The resulting “Cambridge Randomizer” was designed as a web based portal, which was password protected so that access could be restricted to authorised personnel. The portal was able to incorporate both the eligibility screening questions and an embedded random allocation algorithm to be applied to all cases screened in to the experiment. Furthermore, as the portal could email a report of every case, both eligible and ineligible, to the research team, practitioner managers and project staff, it offered the potential for tracking of both exclusions and inclusions without the need for a permanent research office managing in and out of office hours’ random assignments. Despite this, the authors still recommended that there should be an on-site research staff presence in order to enable such functions as “data collection, through experimental oversight, to moral support for the treatment provider staff” (Ariel et al., 2012:201). The Cambridge Randomiser was first deployed in a RCT testing interventions to tackle repeat anti-social behaviour in a UK force. As we shall see below, Operation Turning Point provided a larger and more complex test of the approach.

4.2. (d) Treatment delivery

Ariel et al. (2012) had ultimately concluded that “human contact is terribly important in ongoing research” (p.201). Those conclusions are emphasized in the case studies and commentaries which discuss the process of delivering and tracking treatments. Feder et al. (2000) documented how an ambitious treatment delivery was gradually diluted by middle managers concerned that the pressures created by the demands of the experiment would reflect badly in their performance figures. The researchers had both to “trust” that officers would deliver the original treatment intervention strategy and engage in a formal process of regular progress meetings and informal dialogue to keep the experiment on track. Even then, “there was no doubt that the treatment had become diluted” (p.390).

4.2. (e) A Dedicated delivery team

For Sherman et al. (2015) the experience of conducting the RISE and UK RJ RCTs provided a strong argument in favour of using a dedicated, full time unit to manage the treatment delivery. They contrasted the relative variability in treatment in the RISE experiments, where the delivery teams had shifted throughout the RCT, with the stability provided in the UK experiments by a single specialist team. Berk et al. (1988) and Sherman (2010) both emphasized the importance of recruiting the right staff to such teams and ensuring effective and committed managers.

The decision to opt for a special team rather than seeking to manage implementation within normal routines and deployments creates a potential tension – Eck's diabolical dilemma (2002) – between threats to internal or external validity. Eck described the balance between these two in a RCT as a "trade-off" for evaluators, because the "field settings make it impossible to eliminate these threats entirely". Evaluators could either "be confident of the findings and the lack the confidence that the findings are useful" because the "field setting" had been so artificially manipulated as to make generalisation problematic, or "generalise to many settings, but lack confidence in the findings we want to apply" (Eck, 2002: 104). Eck's "dilemma" may appear to be insoluble but, as we shall see below, it does not take account of the possibility that a highly manipulated intervention, which is evaluated as producing a significant impact, could provide the template for manipulating and reforming operational practice. Defining the right balance between specialist delivery teams or relying on interventions to be delivered within normal operational routines may be just such an example. The specialist team, which may deliver higher internal validity for the experimental conditions may well be better designed, as Sherman et al. (2015) suggest, to deliver high levels of treatment integrity in both experimental and operational settings.

Another way of responding to the "diabolical dilemma is suggested by Gottfredson et al. (2015), who distinguished between the standards required for "efficacy" experiments, in which the emphasis is on testing the hypothesis under highly controlled conditions and "effectiveness" trials, in which there is a greater emphasis on testing the intervention in 'normal' operational conditions. In clinical medical research a similar distinction was made by Schwartz and Lellouche (1967) between "explanatory" and "pragmatic" trials. Patsapoulos (2011) described the former as characterised by high internal validity, the latter by a greater focus on "real world" conditions. This distinction has considerable importance in the field of policing research, where there is evidence of cynicism about police "experiments" that are designed to report success or "foregone conclusions" rather than objective evaluations (Weatheritt, 1986 and Reiner, 2010).

4.2. (f) Tracking

Alongside the human relationship between researchers and practitioners and the quality of the delivery team and its managers, Sherman (2013) and Sherman et al. (2014) have placed a very strong emphasis on the importance of a clear and effective discipline of "tracking" both treatment delivery and dosage. Ariel et al. (2016) argued that the lack of detailed tracking has presented a significant problem in assessing the effectiveness of place based policing, where they argued that only the Sherman and Weisburd (1995) Minneapolis RCT had, until recently, paid sufficient attention to the level and nature of patrol dosages in the treatment and control hotspots. We have seen in the analysis in Chapter 2 that a significant proportion of place based RCTs have provided "unclear" data on treatment delivery.

However, as MacQueen and Bradford (2016) discovered in trying to understand the reasons for the apparent backfire of their ScotCet Legitimacy RCT, tracking can present practical and leadership challenges in policing. They found that the officers conflated the measures taken to track the treatment delivery in the trial with the wider, distrusted performance

management behaviours of managers in an agency undergoing a radical change. As such the research motivation for the trial became confused with the reform rationale, which was being strongly resisted at the frontline. Sherman (1992) described some similar challenges in the Minneapolis Domestic Violence RCT where the Police Chief, who was the champion of the trial, did not enjoy strong support amongst middle management and supervisors. In contrast, in the Trinidad and Tobago RCT, the engagement of regional commanders and their staff was critical to a local "Copstat" process to track and enhance levels of patrol dosage in delivering targeted prevention of gun crimes (Sherman et al., 2014).

4.2. (g) Researchers, practitioners and "ownership of science"

Overall, whether looking at eligibility screening, random assignment or treatment tracking, the case studies and commentaries emphasise the importance of an active and interventionist approach from both researchers and managers. Famega et al. (2016), drawing on Weisburd and Neyroud (2011) and Neyroud and Weisburd (2014a&b), concluded that directing that activity and energy towards achieving "ownership of the science" was the vital determinant of the difference between an increased level of experimental treatment and a level that met a high level of treatment integrity. Merely enhancing the level of tracking and supervision without engaging the frontline in the project and its scientific purpose produced compliance without commitment and, ultimately, a level of hotspot dosage that was insufficient to produce the intended impact.

That sense of needing to marry engineering with vision and values was highlighted by Sherman (2010), who compared the process of conducting and managing a successful RCT to building a cathedral: firstly, it required a clearly defined set of steps - from theory building, identifying the research station, creating a team, pre-testing, getting cases and randomly assigning them, ensuring and tracking treatment delivery and measuring outcomes - that had to be performed in sequence; secondly it demanded practical experimental

skills above and beyond theory and technical methods; but above all, an active and engaged ‘Master-builder’ rather than a passive and independent researcher-evaluator was essential.

The latter role could be seen as creating a conflict between the role of the researcher as an expert facilitator of the experimental process and an independent evaluator (Eisner, 2009). Petrosino and Soydan (2005) demonstrated that there appeared to be a correlation between criminal justice studies with a “developer-evaluator” and higher effect sizes. Eisner (2009) argued that this relationship was a significant enough risk of bias, either through deliberate manipulation of the data and analysis or cognitive bias, that it required a series of measures to minimize the risks. Amongst these were just the type of standards and detailed reporting requirements described by Flay et al (2005) and Gottfredson et al. (2015). However, the real focus of Petrosino and Soydan (2005) and Eisner’s (2009) concerns were on the cases where the researcher had not just an interest in the particular trial, but an ongoing interest in the specific intervention or innovation being tested.

Given that we have identified in Chapters 2 and 3 that there may be a relationship between novice researchers and problems with treatment integrity, maintaining independence should probably be seen as but one dimension of the researcher role. Braga et al. (2014) also identified the importance of a core group of experienced experimenters in conducting police RCTs. Sherman (2010) proposed that the balance of the argument was in favour of experienced experimentalists playing an active role in shaping and testing interventions in such a way as to ensure that they were more likely to be capable of scaling up – a process which Spoth et al. (2013) have identified as being heavily reliant on the quality of the data and implementation advice generated by the field trials. This tension in the role of the researcher is clearly apparent in a number of the juvenile justice trials where apparently bespoke interventions were tested with small samples and never replicated by independent researchers. It is to these studies that we now turn.

4.3 Police-involved juvenile justice RCTs

The 17 juvenile justice police RCTs span more than 45 years (Table 3.2), but the majority were completed before 1980. The key common feature of all 17 is that the core design was a "trickle flow" with cases being entered in to the trials as offenders entered the "gateway" to the criminal justice system (Neyroud, 2015b). As such the trials share many of the issues identified in trickle flow designs discussed above. In particular, screening for eligibility, the method of assignment, monitoring attrition and delivery and tracking of treatments stand out as areas with higher potential risk of bias. We shall look at each of this issues in turn, concluding this section with an analysis of implementation problems in pre-court diversion programmes.

4.3 (a) Eligibility screening

The eligibility screening process in a number of trials appears to present particular risks. In the two trials reported by Davidson et al. (1977) the decisions on eligibility were left to the "discretion of individual juvenile officers" supported by a guideline that simply stated "since the project does not want to become involved with youth who have been involved in only a single minor offense and are not likely to find themselves in further legal difficulty, only refer youth for whom court referral is being seriously considered" (p.42). Given the absence of any tracking data on the population of offenders who were excluded, it is unclear how the sample was screened. This lack of detail on screening and the apparent reliance on police discretion is common to almost all the early trials (Rose and Hamilton, 1970, Stratton, 1975, Binder and Newkirk, 1977, Ku and Blew, 1977 and Lincoln et al., 1977). The consequent risks of bias were examined by Dunford et al. (1982) who reviewed the previous diversion experiments and observed that, in a number of cases, the conclusions would have to be dismissed as "artefacts of the way that comparison groups were selected" (p.7).

4.3. (b) The process and tracking of random assignment

A number of the trials were also silent on the precise method and monitoring of the random assignment process (Stratton, 1975, Binder and Newkirk, 1977, Davidson et al., 1977 and Quay and Love, 1977). Rose and Hamilton (1970) provided the Chief Inspector in charge with a series of sealed envelopes "in which an instruction for caution or supervision, prepared by throwing a dice, had been placed". Their demographic data comparing the treatment and control groups suggests that there was a significant difference between the two groups with more serious offenders in the control group. The sealed envelope method was, where disclosed, the most common method deployed. Given the very significant risk of bias from this method that we have set out above, it is concerning that only McGarrell and Kroovand (2007) describe detailed tracking of the consistency of the subsequent random assignment by the researchers.

Byles and Maurice (1979) adopted a different method of random assignment and reported tracking it. They used the odd and even case numbering produced by the agency's internal file recording system, which they claimed could not be manipulated. However, their analysis of the characteristics of the sample showed a statistically significant difference between the treatment and control samples. The "E", experimental group, had more police involvement but significantly fewer serious cases, suggesting that the random assignment method was far from fool proof. Indeed, just as Sherman and Berk (1984) experienced officers using their discretion to divert more serious offenders into the arrest treatment, it is quite apparent from Byles and Maurice (1979) and a number of other juvenile justice trials (Dunford et al., 1982 and Little et al., 2004) that the risk that officers will remove "difficult cases" from diversion presents a serious risk of bias in pre-court diversion experiments. In support of this, Lincoln et al. (1977:185) and Klein (1986) reported unequivocal evidence that the cases were deliberately manipulated in this way by officers. In the only post-conviction police involved study, Little et al. (2004) reported that "practitioners managed to by-pass procedures" and the court acerbated this position by ignoring random assignment in two further cases (p.231).

Dunford et al. (1982) encountered difficulties all through their multi-site, multi-agency diversion trial. However, it is notable that the most serious difficulties with attrition were experienced when the field researcher who was supposed to have been tracking the cases entering the trial in one of the field sites went absent without leave for an extended period. The hiatus allowed both cases and data to go missing and presented a serious risk of bias. Furthermore, in all four RCT sites, Dunford et al. (1982) encountered problems of reassignment from control to treatment and within treatments by police, probation and justice officials who continued to exercise their discretion despite the experimental protocol. They did so to such an extent that Dunford et al. observed that "because some youths in this study did not receive the services to which they were assigned, and because some youths received services from dispositions to which they were not assigned, this research design did not represent a pure test of diversions programs" (1982:29). They went on to suggest, drawing on their review of the previous evaluations, that this may well reflect the problems of implementing diversion schemes in "normal settings". In that sense, the trial, whilst largely unsuccessful as a RCT, was still, in this view, a credible case study of the operational challenges of pre- court diversion.

4.3. (c) Treatment dosage and delivery

However, such a judgment would be strengthened if the evaluations had been more specific and detailed on the treatments and dosage actually implemented. Quay (1977), drawing on lessons from his own trial (Quay and Love, 1977) had identified the critical importance for "program integrity" of specifying the hypothesis and treatment design, the tracking and monitoring, duration and intensity of treatments. Yet, the level of detail provided on the nature, dosage and tracking of treatments varies widely. Most of the trials do not, as Flay et al. (2005) and Gottfredson et al. (2015) exhort, provide such details as would enable an accurate replication. Rose and Hamilton (1970) only state that the treatment group were "cautioned and supervised" for up to

six months. We are provided with no details on attendance, level of dosage or the nature of the interaction. In a similar vein, Lincoln et al. (1977) and Klein (1986) tested four treatments but provide such limited data that it is not possible to understand the duration, intensity or compliance. Stratton (1975), Davidson et al. (1977), Ku and Blew (1977), Quay and Love (1977), Byles and Maurice (1979) and Giblin (2002) all provide more detail of the proposed treatments, but only Byles and Maurice (1979), Giblin (2002) and Little et al. (2004) give a breakdown of completion and some indication of dosage. On the other hand, the small samples and nature of the intensive supervision documented in studies such as Ku and Blew (1977) suggest that the core mentoring and supervision model was delivered.

Even with the limitations of the published data, it is quite apparent that there were some significant treatment failures: in Byles and Maurice (1979) more than half the sample refused the intended treatment; Giblin (2002) reported that less than half the sample were provided the treatment, admitting at the same time that the trial did not truly amount to a test of the intended treatment. Indeed, Giblin's experiment appears to be an example of an unsuccessful effort to maintain the "coalition" of partners in order to secure consistent treatment through the trial period (Strang, 2012). In a similar multi-agency programme, Little et al. (2004), despite a small sample, only managed to ensure an average of 64% of the required treatment elements were delivered to the 24 treatment cases in the required dosage.

4.3. (d) Pre-court diversion and the problem of implementation

Given the generally low level of treatment integrity in these trials, it is worth returning to Dunford et al.'s (1982) suggestion that there are inherent factors in the nature of pre-court diversion that may create higher risks of attrition and performance bias. UK evaluations of pre-court diversion suggest that there may well be some such consistent themes. In a succession of studies of police cautioning, Mott (1983), Laycock and Tarling (1985) Giller and Tutt (1987), Sandars (1988) and Evans and Wilkinson (1990) found a wide and

apparently unjustifiable diversity of practice and concluded that this affected the fairness and justice of the system. This can partly be explained by the need, as identified by Schwalbe et al. (2012)'s meta-analysis of pre-court diversion, to "tailor" the approach to the individual needs of the offender. However, as Landau and Nathan (1983) found, there are also other factors including prejudicial judgments about offenders based on their race or class.

There were clearly factors beyond "tailoring" at work in accounting for the relative failure of a succession of initiatives to introduce individualized treatments – conditional cautions – and triage schemes to identify the appropriate offenders and treatments. Blakeborough and Pierpont (2007) conducted an evaluation of the pilot implementation of Conditional Cautioning for the Home Office. They found considerable variation between and within the six forces studied. Despite the fact that the scheme was intended to deal with a wide range of offences and introduce a greater use of restorative justice, more than half the conditional cautions were for minor criminal damage with the most frequent condition set being financial compensation to the victim. Reviewing the data on the deployment of conditional cautions over the decade since their introduction Neyroud and Slothower (2015) found that this pattern of divergence from the original policy objectives in conditional cautions has remained consistent.

Two further studies examined the introduction of specific conditions for women (Easton et al., 2010) and community payback conditions (Rice, 2010). Both studies found problems in the custody suite with custody officers finding the conditions complex and the task of matching suitable cases to the conditions one that they were unprepared for by training or guidance. Taken together these studies suggest that it may be hard to shift the operating model of the police towards a more restorative and rehabilitative model and away from a presumption to prosecution or crime control model, particularly where, as in the UK, such embedded assumptions have been reinforced by performance management systems (Cockcroft and Beattie, 2009).

The lessons from the evaluations of police custody triage and Liaison and Diversion programs (ICPR, 2012, Moore et al., 2012, Wood et al., 2011 and Haines et al., 2012, Ames et al., forthcoming and Disley et al., 2016) suggest that the challenges of shifting the cultural model of policing may be intertwined with practical and logistical problems with implementation. The triage programs all involved dedicated staff – some Youth workers, some police officers and some third sector staff – in assessing young defendants brought to the custody suite. The primary aim of the programs was to identify defendants suitable for diversion and shape a programme of interventions. ICPR (2012) found that the high turnover of custody staff, their lack of awareness of the program and the lack of feedback to them about outcomes were all serious problems. Haines et al. (2012) also found that only a third of defendants took up the triage and treatment on offer. Others were assessed, but the slippage as result of staff availability and communication between custody staff and case workers accounted for a larger group than those treated. The RAND study of Mental Health Liaison and Diversion schemes was generally more positive. After initial difficulties in recruiting, training and vetting staff, which meant that effective implementation took more than a year, the dedicated teams seem to have identified, diverted and treated significant numbers (Disley et al., 2016). The experience of the triage and Liaison and Diversion schemes appears to lend weight to Sherman's (2010) contention that small, dedicated teams may be required in this field to ensure consistent delivery and treatment.

This overview of juvenile justice RCTs and studies of pre-court diversion in the UK would seem to suggest that this is a particularly challenging area both in service delivery and to test experimentally. A recent process evaluation of a 3 force pilot of a new approach to Out of Court Disposals provides further confirmation of those challenges (Ames et al., forthcoming). The aim of the pilots was to test a "simplified" version of the existing out of court disposal regime, the key feature being a reduction from seven possible disposals to two – a community resolution plus and a conditional caution plus. The pilot and evaluation took place in 3 forces with 3 further forces identified to provide control data.

The evaluation identified serious implementation problems. In reflecting on these, the final report highlighted the need for more careful preparation, pre-testing before go-live, a more effective face to face training approach and more assistance and support to officers on setting conditions. Alongside this was a need for greater recognition of the substantial role change that conditional pre-court disposals required from the police. The shift from deterrent enforcement to preventive rehabilitation was not seen as unwelcome by the officers surveyed, but the evaluation did suggest that it moved them into areas outside their training and traditional skills set. As one senior officer observed “the culture of policing is interview, charge. Interview, caution. Interview, bail: the old traditional outcomes. It is a challenge to encourage police officers to suddenly start thinking differently about what appropriate conditions and interventions might be attached to a caution” (Ames et al., forthcoming: 4.2)

As a result, the limited preparation and training for the pilots – a short session supported by online material for most staff – was seen as entirely insufficient. Indeed, because of the short run in to the live phase of the programme, all the forces in the pilot identified that some key treatments had not been available when the study evaluation started. Overall, Ames et al. (forthcoming) highlighted that pre-test preparation, investment in training and skills, management and leadership engagement and interventions and tracking delivery were all critical in delivering a pre-court intervention and had only been partly delivered in the MoJ sponsored pilots (Ames et al., forthcoming).

4.4 A framework for treatment integrity in Police RCTs

Drawing on the analysis in Chapters 2 and 3, the cases studies and juvenile justice RCTs discussed above in this chapter, we can begin to outline a framework of key issues that researchers and practitioners need to pay attention to in order to address treatment integrity in the design and implementation of experiments, particularly where the intervention either

centres on a pre-court disposal or requires a trickle-flow design. Following the framework that we used in Chapter 2, Table 4.1 is divided into issues connected with managing attrition and issues that appear to be important for securing high levels of treatment delivery.

Attrition	Performance
• Pre-Test: resolving legal and ethical issues - Dunford (1990) • Pre-Test: dry run of random assignment process – Dunford (1990) • Design and testing of the eligibility screening process – Earle (1973) and Dennis (1988) • Random Assignment method – Braucht and Reichardt (1993), Sherman (2010) and Ariel et al. (2012) • Designing overt overrides and securing ownership – Dennis (1988) • Blinding of practitioners to the Random assignment – Sherman (2010) • Separation of operational staff from control of the random assignment process – Sherman (1992) and Sherman (2010) • Tracking of random assignments – including the pre-assignment screening and post-assignment attrition	• Pre-test: dry run of the treatment approach – Sherman (2010) • Treatment Protocol – Gottfredson et al. (2015) and Design of the treatment. • The “treatment team” – dedicated or within “normal” operational roles – Sherman (2010) • Training and briefing (Ames et al., forthcoming) • Process for securing consent – Sherman et al. (2015) • Field Support – Dunford et al. (1982) • Tracking of treatment delivery, duration and dosage – Quay (1977), Sherman (2010) and Sherman et al. (2014)

– Sherman (1992) and Sherman (2010) • Tracking the "hydraulic flow" through the experiment – Strang (2012)	

Table 4.1: Factors affecting treatment integrity in "trickle-flow" police RCTs

Sherman's (2010) incorporated many of these issues into his building blocks of successful experiments under four headings: screening for eligibility; assigning treatments; delivering treatments consistently; measuring treatments delivered. After describing the Turning Point case study, the background, hypothesis, initiation and pre-testing, Sherman's four headings and the key issues in Table 4.1 will be used to structure the analysis and discussion.

4.5 The Turning Point Case Study

4.5.1 Introduction

The case study has been developed by the author whilst conducting the Turning Point experiment as the research manager and one of the principal investigators. The data gathering for the case study was part of the ongoing evaluation of the experiment and has been carried out in parallel with the author's research on Police RCTs which has been set out in Chapters 2 and 3. Yin (2014) recommends using multiple sources of evidence in a case study

in order to provide triangulation on the same research question. The primary sources of evidence for the case study fall into three parts: firstly, documentary evidence from the minutes of meetings, emails and documents which were generated by the process of the experiment, which have been edited and catalogued as “DOC/TP/number” for reference (Appendix 6); secondly, the records generated by the Cambridge Gateway randomizer, the treatment records and the electronic treatment portal; lastly, personal observations from visits to police stations and custody suites and from training events, seminars and conferences at which Turning Point was discussed or reported.

In the classic case study method described by Yin (2014), the sources are chosen by reference to a research question and designed to converge on the findings. In this case, the original research program was tightly focused on the experiment and its evaluation. The sources were gathered initially as part of the evaluation rather than purposively as part of an exploration of treatment integrity. However, as the experiment progressed, the researchers encountered significant problems of implementation and treatment integrity. As will be set out below, these were so severe by the middle of Stage 3 that we were forced to take the unusual step of restarting the experiment as a direct result of attrition and performance problems. This researcher, like Dunford (1990), Sherman (1992), Feder et al. (2000), Famega et al. (2016) and MacQueen and Bradford (2016), recognized that the Turning Point experiment was capable of providing important insights on the conduct and management of experiments in policing. The approach adopted to the material has, therefore, been largely deductive, drawing on the search for and analysis of Police RCTs, as set out in Chapters 2 and 3 and the analysis of existing case studies and juvenile justice RCTs as summarised in Table 4.1 above.

Although the three sources of evidence are the main reference points for this chapter, the research team also conducted 18 semi-structured interviews with a purposively selected sample of staff from custody, offender management teams, program team and senior management. These interviews were framed

against a protocol which was designed to address the wider issues of conduct and management of the experiment and the process of learning from RCTs as part of collaboration between this researcher and a colleague at University of Queensland (Bedford and Mazerolle, 2014). The protocol (Appendix 3) was designed by this researcher and the interviews delivered by Bedford. The approach to the interviews and their analysis will be set out and discussed in more detail below in Chapter 5 in which we turn from the detailed questions of treatment integrity to the broader framework of conduct and management of experiments in policing. The interviews were open coded using HyperResearch for Mac and a number of the categories relating to randomization, treatment delivery and tracking of the experiment are relevant in this Chapter as well.

4.5.2: The method, role of the researcher and the status of the research

The process of developing the case study and managing the experiment raised some important, not to say unique, issues around the role of the researcher. We have seen in Chapter 2 how the analysis of police experiments has demonstrated an increasing involvement by practitioners not simply in research partnerships engaged in Police RCTs, but also as researchers. Shepherd (2003) noted that "Evaluation of medical interventions has traditionally been led by practitioner (clinical) academics. This is not the case in criminal justice, where theory has had higher status than intervention research." (p.290). Shepherd argued for more practitioner led research, more practitioner leadership of research and for the development of a model in policing that could be compared to the teaching hospital.

This author drew on Shepherd's proposal in recommending the creation of a Professional body for policing and the embedding of evidence-based approaches as core themes of the Review of Police Leadership and Training (Neyroud, 2011b). Whilst completing that Review, the author started the early work on the Turning Point experiment whilst still a serving Chief Constable in 2010, in a role, as Chief Executive of the National Policing Improvement

Agency which included strategic and operational responsibility for police research across England and Wales. Prior to that, during a professional career spanning more than 30 years, the author was involved with research, policy development and the implementation of policies and practices on pre-court diversion at a number of key points. These included: as a young constable working with Giller on an evaluation of instant cautions in Hampshire (Hampshire Constabulary, 1982); as an Inspector, a scrutiny of pre-court diversion in Hampshire and design of a new approach (Hampshire Constabulary, 1988); as a Chief Constable, the idea for and the piloting of conditional cautions, including responsibility for the drafting of legislation and the development and delivery of the programme to implement Conditional Cautions (CPS, 2016); As the National Chief Officer lead for Out of Court Disposals, the creation of the gravity matrix for cautioning and prosecution (ACPO, 2009) and, as a member of the Sentencing Guidelines Council (SGC), their endorsement by the SGC. The author was subsequently the co-author of a "controversial" (Engel and Henderson, 2014) set of papers advocating the transformation of the police through science (Weisburd and Neyroud, 2011, Neyroud and Weisburd, 2014a and 2014b), which argued for greater police ownership of and engagement in the science of policing.

In summary, the author was, whilst serving as a police officer, one of the leading professional contributors to the development of the existing legislation and practice on out of court disposals in England and Wales. On retirement, whilst conducting and managing the experiment, the author had a number of roles: the principal investigator for the RCT; the research manager responsible for managing the research programme; a PhD student; a part time, affiliated, and then full time Lecturer responsible for teaching experimental criminology to a number of those involved with the research program who had become Masters students; a member and subsequently Co-Chair of the Campbell Collaboration Steering on Crime and Justice; an independent expert for the Ministry of Justice evaluation of the "simplified" model of OOCD (Ames et al., forthcoming) and for the RAND study of Liaison and Diversion (Disley et al., 2016).

Whilst the author – part police officer, part academic – as one of the principal investigators was a “novice” RCT researcher, as was the field researcher, the research was supervised by Sherman, one of the most experienced “super” experts on police RCTs identified by Braga et al. (2014) and supported by another, Strang. Ariel, who had already been a principal investigator on the first West Midlands police RCT, was also involved in the research design and the development of the Gateway. The author was, therefore, operating in an environment – the police station – which was familiar, but in a role and context which was novel but with the support and supervision of experienced experimental research investigators.

In reflecting how such different roles might matter in the research process in policing, Brown (1996) suggested that research in policing could be described by a typology framed by the identity of the researcher: “inside-insider” researchers tended to conduct “official police research” (Reiner, 2010) for the organization; “outside-insider” research was conducted by former members of the force who had migrated towards academic careers; “Inside-outsider” researchers either work in dedicated research units within forces or are embedded in the force; “outside-outsider” research covers most academic research on the police. This research, following Brown’s typology, could be said to have started as “inside-insider” but has evolved into outside-insider research as the researcher has moved further from policing and closer to an academic role.

However, in many ways Brown’s typology seems over-simplistic to frame the Turning Point project and a growing number of practitioner-led or practitioner involved RCTs in policing (Appendix 2). Sherman (2009) began to outline a complementary framework to Brown’s typology. In his proposals for advancing experimental criminology, Sherman discussed the relative merits of “bottom-up” discretion versus “top-down” guidance. In his analysis, the first, informed by practitioner access to tools such as systematic reviews, offered the opportunity to build better evidence from innovative practice, the second, utilising the resources of the state and its agencies, the possibility of framing the research agenda and guiding investment in the most promising areas.

Underlying Sherman's approach were four approaches to evaluating innovation and practice (Sherman, personal communication, 2015):

- **Innovation without evaluation**: an example of this would be the development of DNA testing for forensic investigation. Whilst the science of DNA was tested (National Research Council, 2009), the effectiveness of DNA as a means of investigating crime had to wait for the Urban Institute report (Roman et al., 2009), more than 20 years after the first use of DNA in crime investigation in the UK in 1987. DNA effectiveness was, initially, therefore, consistent with Brown's "insider-insider" model, with reliance placed on operational and official data rather than experimental designs.
- **Academic top-down experimentation**: as Weisburd and Neyroud (2011) outlined and the analysis in Chapter 1 has demonstrated, this was the dominant model of research evaluation in policing for the first 40 years of experimentation. This approach most clearly typifies Brown's "outsider-outsider" research model.
- **Practitioner-led experimentation**: Although Earle (1973) provided a very early example of this, Farrar and Ariel's Rialto Body Worn Video RCT provides the most influential example, which has already stimulated a further fifteen RCTs and other studies seeking to replicate or explore wider aspects of the use of the technology. This model of research does not easily fit Brown's typology. The closest match is "outsider-insider" research because, in most cases, the practitioner, such as Farrar, was temporarily "outside" the organisation as a student, subject the requirements of a Master's degree and academic supervision, when the research was carried out.
- **Down and Up Experimentation**: In this last, academic thinking and practitioner innovation combine to produce a project driven locally, in partnership with the local force, but which has the potential to drive the wider national or international evidence agenda. "Down and Up" experimentation does not easily fit in to Brown's typology, with neither

> "outside" nor "inside" in the lead but rather combining elements of both to produce the outcome.

The researcher's role and the nature and status of the research are, as both Brown (1996) and Sherman (2009) agreed, important in considering the status, potential bias and independence of the research findings. Yin (2014) identified that case study research requires a combination of skills, including the ability to ask good questions and a firm grasp of the issues being studied, but that these need to be balanced by objective listening, the avoidance of bias and ethical research practice. Nutley et al. (2007) argued that insider insights and professional knowledge can be critical to framing and interpreting research. Yet, until recently, practitioner expertise and the researcher were almost always separated in policing (Bradley and Nixon, 2009). Shepherd (2003), in comparing medical and policing contexts, observed both the lack of police practitioner-academics and the fact that this limited the motivation and process of research to Brown's (1996) "outside-outsiders". In particular, for Shepherd, this meant that it was more difficult to translate an observed problem in police practice into a research programme to test an alternative practice.

The focus of this case study, Operation Turning Point, was a complex experiment designed to test an alternative model of pre-court disposal. The force under study was committed to an evidence-based approach and one of the local commanders was a leader of the emerging Society for Evidence-based policing. On both sides of the partnership, police and research team, there was an overt commitment to scientific testing of practice innovation and a shared vision of the police station as a learning and teaching environment as set out by Shepherd (2003). This author has suggested (Neyroud, 2016), in writing about the ethics of practitioner-led police research, that Turning Point, therefore, offered insights on both the strengths and limitations of police ownership of science. An "insider" view could be seen as biasing the data gathering and reporting of outcomes. On the other hand, Independent funding, a commitment to peer reviewed publication of the outcomes and

University supervised field research have provided an important counter balance to a professionally led and managed program. Whilst in Hirschi's terms, the research would undoubtedly qualify as "administrative criminology", for Shepherd (2003), approaching the issues from a medical framework, a better label is "applied criminology" contributing to a strategy in which "research, practitioner training and services" are integrated (Shepherd, 2015:3).

As will become clear in the case study, Turning Point followed very much this model, with a research effort led by a hybrid team of academics and practitioners, a constant attention to developing evidence-based knowledge and practice in the force and a commitment to design and scale up from the research to a new national approach to service provision in pre-court diversion. As such, Turning Point would appear to match most closely to Sherman's "Up and Down Experimentation" model as well as providing an important case study for Shepherd's model of practitioner education and research (Shepherd, 2003).

4.6: The Turning Point Trial

4.6.1. Overview of the experiment:

Operation Turning Point was a randomised controlled trial (RCT) designed to test whether low harm offenders, both adult and juvenile, who might otherwise have been prosecuted, could be effectively dealt with by a combination of deferred prosecution and tailored conditions to encourage desistance. The trial was conducted in four Local Policing Areas within the West Midlands Police Area: the original two areas, Birmingham East and Birmingham South; two further areas, Birmingham North and Birmingham West and Central. The final area covered by the trial included the whole of the city of Birmingham (population 1.085 million[3]).

[3] Birmingham City Council and ONS estimates: http://www.birmingham.gov.uk

The trial involved taking offenders whom the police have decided to prosecute, screening out those with more than one conviction or a recent conviction and those excluded by a set of qualifying questions. Those offenders screened in as eligible were randomly assigned to either a control condition, prosecution, or a “turning point” treatment. This centred on a “contract” negotiated between the offender and a police officer from the Integrated Offender Management team or a member of the Youth Offending Service (depending on the age of the offender). Compliance with the contract over a period of four months meant that the threatened prosecution was dropped. The contracts included two key elements: a requirement not to reoffend; a requirement to comply with the tailored conditions that were intended to be matched to the offender’s pathways to crime.

The allocation was carried out automatically by a computer based randomiser – the Cambridge Gateway, which was set to allocate eligible cases randomly on a 50-50 basis between control and treatment. The Gateway was operated by police custody officers, who were required to use it every time that an offender was deemed to have met the threshold test for court prosecution – sufficient evidence for a realistic prospect of conviction and in the public interest to prosecute (Moreno and Hughes, 2012).

The experiment was conducted in four phases:

- Stage 1 (November-December 2011) involved pre-testing the Gateway and the eligibility screening;
- Stage 2 (December 2011-May 2012) centred on pre-testing the treatment delivery;
- Stage 3 (June 2012 to April 2013) was the first phase with full experimental conditions;
- Stage 4, which ran from April 2013 to July 2014) was, effectively, a restarting of the experiment. It also involved two experiments embedded within the same trial – the test of the diversion and a test of the victim experience between diversion and prosecution.

In Stage 4, the Gateway was amended to block randomise the offenders in four conditions: adult offender with a personal victim; adult offender with no personal victim; juvenile offender with a personal victim; juvenile offender with no personal victim. The amendment allowed the research team to create a randomised sample of victims who could be interviewed about their experience and satisfaction with the process (Slothower, 2014). Additionally, the block design recognised the fact that two distinct teams were treating the juvenile and adult offenders – the Youth Offending service for juveniles and offender managers for adults.

There was a significant contrast in the level of attrition and treatment delivery between Stage 3 and Stage 4 as set out in the CONSORT diagrams for the experiment (Figures 4.2 and Figure 4.3). Stage 3, which had been intended to be the main data gathering phase for the final evaluation, had to be curtailed, because the research team found substantial attrition from assignment and issues with the quality of treatment specificity in the contracts being set. The overall treatment integrity in Stage 3 was less than 60%. By contrast, in Stage 4, that figure rose beyond 90%. As such Turning Point provides a number of key features that make it particularly relevant as a case study: alongside a small subset of police experiments (Earle, 1973 and Weisburd and Green, 1995), Turning Point was paused and effectively restarted, but with changes to the design and controls; like those two experiments, the restarted experiment produced considerably higher levels of integrity, suggesting that a comparison between the two stages may provide significant evidence on the factors contributing to a successful, high integrity experiment.

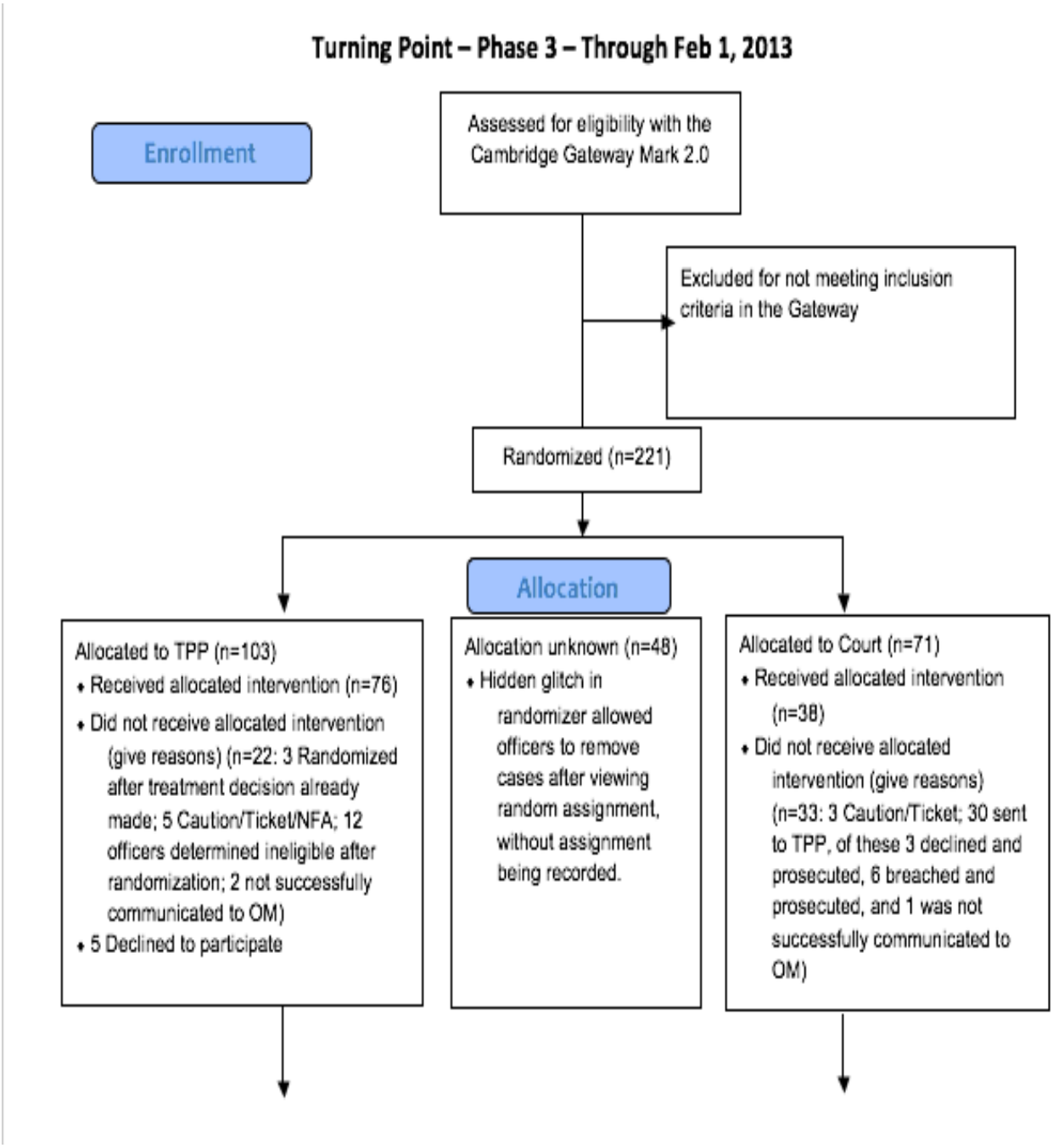


Figure 4.2: CONSORT diagram of Stage 3 of the Turning Point Experiment.

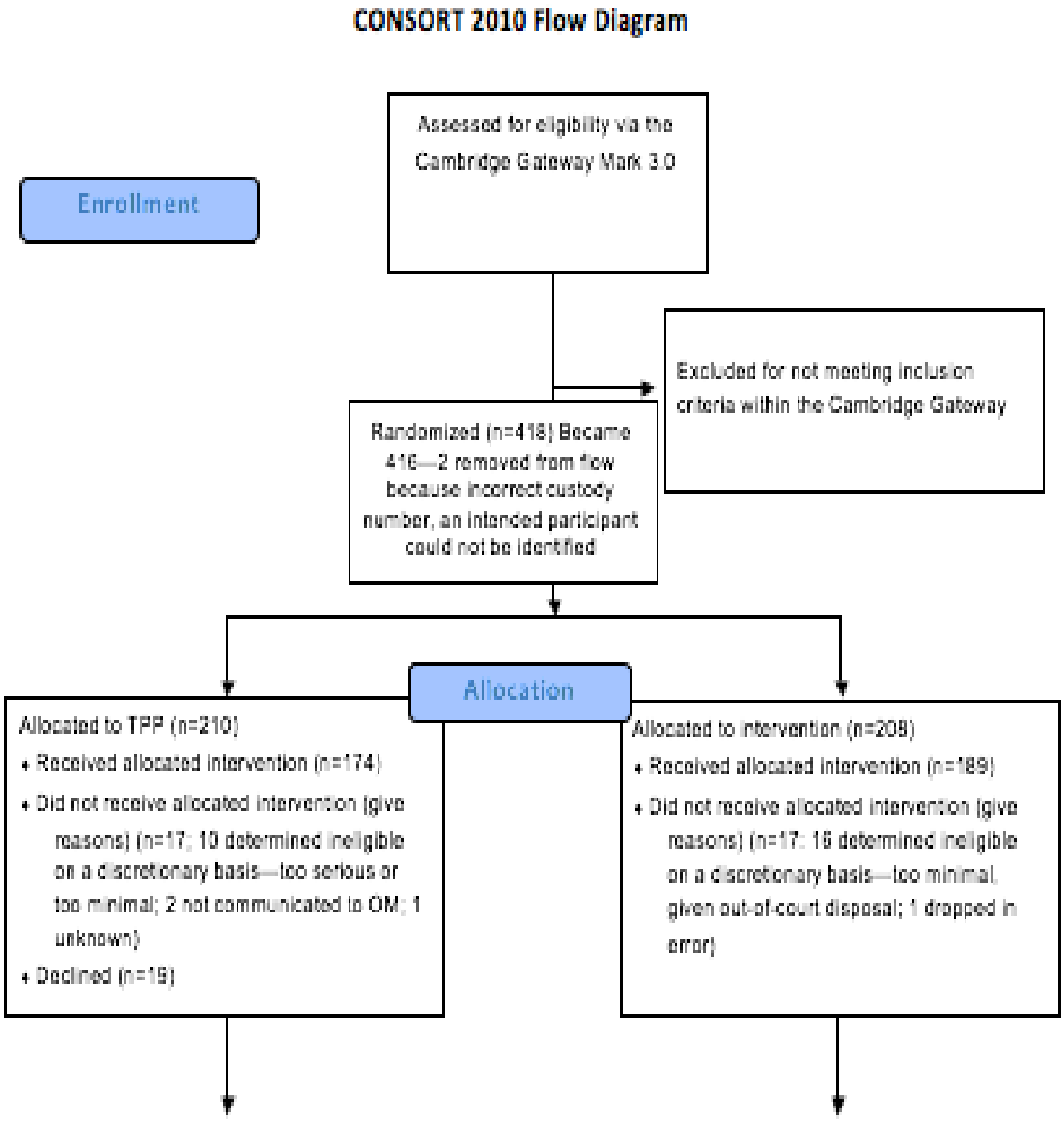


Figure 4.3: CONSORT diagram for Stage 4 of the Turning Point Experiment

4.6.2. Preparation and Planning for the Experiment:

4.6.2 (a) Origins of the Experiment and Research Motivation:

In his inaugural lecture as Wolfson Professor of Criminology, Sherman argued that experimental criminology could be harnessed to enhance liberty by

identifying and testing key questions “affecting life and liberty” (Sherman, 2009:22). One such area of criminal policy was the process of prosecution, which offered opportunities for the police and other agencies to “effectively intervene to prevent offending, deter offenders and encourage them to desist from further offending” (Sherman and Neyroud, 2012: 2). In developing alternative hypotheses, Sherman drew on the lessons from research on crime-harm forecasting of high harm offenders (Berk et al., 2009), deterrence (Dunford et al., 1990, Durlauf and Nagin, 2011 and Hawken and Kleiman, 2011), desistance (Laub and Sampson, 2001 and 2003), parole management of low and high harm offenders (Barnes et al., 2010) and the negative consequences of formal processing of (juvenile) offenders (Petrosino et al., 2010). He argued for a new discipline of “Offender-Desistance Policing” (ODP) (Sherman, 2011a&b) in which the police would use crime harm forecasting in order to triage their approach to offenders into two broad approaches: an intensive offender management of those predicted to have a high probability of committing a high harm crime; low intensity, preferably pre-court intervention with low harm offenders, designed to encourage desistance.

In England and Wales, the development Sherman’s thinking on the development of ODP came as a new Coalition Government imposed deep cuts on police and criminal justice budgets (Sherman and Neyroud, 2012). Within this context, ODP was framed as an “efficient” approach to criminal justice reform (DOC/TP/1). Furthermore, it coincided with an emerging agenda for reform of pre-court disposals with a view to enhancing their transparency and effectiveness. The Criminal Justice Joint Inspectorate (CJJI) completed a review of “Out of Court Disposals” (OOCD) in 2011 (CJJI, 2011) which had suggested that OOCD’s were probably an effective way to deal with offenders, compared to prosecution, but that the existing regime was bureaucratic, inconsistent and in need of reform. However, although it was supported by a Policy Exchange study (Sosa, 2010), CJJI’s analysis was based on a very small sample of case that were selected by “dip sample” rather than any robust form of random sampling.

In this context, Sherman and Neyroud (2011 and 2012) proposed a research programme to develop and test ODP which would be focused on two areas: the creation of a crime harm forecasting tool using data from the Police National Computer; "first offender experiments" in which first time offenders would be offered an opportunity to work on their offending with a "Damocles squad" under threat of a deferred prosecution. The programme secured independent funding from the Monument Trust and the researchers went in search of police forces prepared to consider becoming the research station for the experiments.

The final choice of West Midlands Police as the host force was driven by a number of considerations: the force had an existing experimental research partnership with Cambridge University in which a RCT to test interventions to tackle repeat anti-social behaviour was already in the field; as the second largest force in the country, West Midlands Police offered both scale and credibility to any experiment, which would be more difficult to achieve in a smaller force; the Chief Constable was supportive and had a strong, personal relationship with both the author[4] and Sherman.

West Midlands Police and Cambridge University agreed in principle to partner to undertake the ODP research programme as set out by Sherman and Neyroud (2011). A joint programme board was created and met for the first time in May 2011 (DOC/TP/2), at which the author presented the research proposal (DOC/TP/1). The proposal had envisaged a "first offender" programme in which first time offenders whose case was deemed suitable for prosecution would be referred to a supervisor and "on site member of research staff". Consistent with Sherman's (2010) advocacy of a separation of practitioner and research staff functions, the researcher would then be responsible for managing the randomisation process. Those offenders randomised to the treatment intervention were to be referred to the

[4] the Chief Constable had been the author's Deputy in the National Policing Improvement Agency.

““Damocles Squad”, or whatever the force chooses to call it” (DOC/TP/1:11) and the treatment set out in Figure 4.4.

1. Meet with the arrestee to discuss the instant offence and its implications for his future.
2. Assess the offender's assets for informal social control, including family, education, employment and community organizations.
3. Consider a range of tactical options for the set of conditions the Damocles Unit will offer the arrestee as a voluntary alternative to prosecution, conditions that may well be in the public interest to prefer over attempts at prosecution.
4. Offer the arrestee the chance to ask questions about the offer, and to consult with a solicitor if the arrestee requests that opportunity.
5. If the arrestee refuses the offer, then the case would be referred for prosecution.
6. If the arrestee accepts the offer, the "Damocles Unit" would ask the offender to sign the statement of conditions, which would not need to include any admission of guilt for the instant arrest offence. The signature would merely indicate the arrestee's agreement to comply with the conditions, and the police agency's agreement not to prosecute the arrestee on that offence as long as the arrestee complies and does not reoffend. The exact content of the conditions is discussed below in the section on "specific tactics.".
7. Police would then follow up to insure compliance with the conditions, basing resources invested on the subjective level of risk the Damocles Unit sees in leaving the arrestee unmonitored. Early monitoring, in any case, could be more intensive, then tapering off as long as the arrestee complies.

Figure 4.4: Treatment proposal for "Project Damocles": excerpt from DOC/TP/1.

The debates at the first meeting of "Project Damocles" centred on a number of key aspects of the research proposal. Firstly, there was the question of whether offenders should be required to admit guilt before being deemed eligible. For a caution or conditional caution, an admission was a legal requirement, but offenders in the proposed trial were going to be subject to a deferred prosecution rather than a caution. Secondly, there was the process for randomisation and referral for treatment. In both the first and second meeting (DOC/TP/3) the tension between conducting the trial within "normal operations" and trying to restrict the number of staff involved was a major focus. However, overriding both was a concern to ensure that the experiment was legal and ethical.

4.6.2 (b) Legal and Ethical considerations:

Dunford (1990), drawing on the experience of conducting the National Diversion programme evaluation (Dunford et al., 1982), argued that resolving the legal and ethical issues in a proposed trial was a potentially critical success factor. He had found that the central concern of the practitioners was on the perceived inequity and injustice of randomly assigning subjects with similar circumstances to two different treatments in the criminal justice system. Kerr et al. (2011) found similar clinical, legal and practical challenges in evaluating the feasibility of a potential post-conviction RCT of treatments in the UK. In their study they found that local practitioners in an England and Wales context were keen to have "top cover" from national officials.

In Operation Turning Point the researchers had set up a small national advisory group linked to the Monument Trust grant. That group included, amongst others, a senior Judge, a senior Crown Prosecutor, the Chief Constable with the national responsibility for OOCD policy and the head of the National Offender Management Service. On a number of key questions, such as the overall legality of random assignment in the prosecution process, the "Project Damocles" board was able to draw comfort from the support of the national advisory board. Furthermore, through the trial the advisory board members were briefed by the research team on the progress of the trial. Their support appeared to provide the necessary level of "top cover" to local prosecutors and probation staff.

It is a feature of police RCTs that professional and operational concerns overlap with the ethical requirements which researchers need to consider (Neyroud, 2017). Punch (1986) and Miles and Huberman (1994) summarised these as: the worthiness of the project; the question of informed consent; the benefits of the research; the potential for harm and risk; the integrity of the research and the ownership of the data and conclusions.

4.6.2. (c) Managing Risk and Harm

For the project board, managing harm and risk and the potential threat to the organisation's reputation stood out as the key considerations (DOC/TP/3). The project board members were, therefore, concerned to ensure that the design of the eligibility criteria excluded higher risk offenders as far as possible. This led to decisions to exclude cases where an offender had convictions or was likely to receive a custodial sentence or was in custody for offences involving knives or sexual offences. Minimising organisational risks dictated that hate crime and domestic violence cases, which were seen as controversial areas for diversion, were also excluded (DOC/TP/3). The decision was reinforced by a national policy debate about the appropriateness of diversion in domestic violence cases, which led to the the Home Office Circular 16/2008 for Simple Cautioning being replaced by new guidance which strengthened the advice against cautions for domestic violence.

As to the worthiness and benefits of the research, Dunford's team had also identified that the criminal justice agencies needed a justifiable narrative, both for internal and external consumption (Dunford, 1990). In the National Diversion programme, the narrative that was developed, after discussion between the researchers and practitioners included three dimensions: the proposed intervention was likely to be effective; no subject would be worse off as a result of the trial; the question – the relative effectiveness of diversion – was both important and, as yet, unanswered.

4.6.2. (d) The Narrative

The narrative that was developed in the Damocles project board was somewhat similar. The research proposal (DOC/TP/1) was seen as providing sufficient evidence to support the testing of the ODP hypothesis and justify the "worthiness" of the research. The board agreed a procedure designed to ensure that there was compliance with the fairness principle that offenders should not be worse off as a result of randomisation, nor should the treatment provide a disproportionately tough set of conditions compared with the trial nor should they be assigned to a treatment that caused them harm or worse

consequences than the control (Torgerson and Torgerson, 2008 and Edwards et al., 1998). In order to achieve these conditions, offenders were not to be put into the Gateway until and unless a clear decision to prosecute had been taken by the police custody officer. Thus, save for the random assignment to the treatment, they would have been prosecuted and be at risk of a criminal conviction and punishment. By being offered the Turning Point trial, their prosecution was deferred. The treatment conditions were designed, as far as possible, by drawing on the research evidence so as to ensure that conditions offered were supported by the best evidence available and that treatments identified as causing increased levels of offending were avoided (Neyroud and Slothower, 2012). Finally, the accuracy of the eligibility decisions and the proportionality of the treatment conditions compared to the potential sentence in court was to be independently audited by the Crown Prosecution Service, who were to be asked to examine a sample of the early cases.

4.6.2. (e) Consent

The process for obtaining consent from offenders to participate in the experiment posed a more difficult challenge in aligning operational considerations, research design and research ethics. It was decided that, given that the decision to prosecute was to be taken before the random assignment, offenders assigned to Turning Point would not be asked for consent until after assignment and would only be asked to consent to detailed conditions after their interview with offender manager. This approach was dictated by the reluctance of the project board members to impose additional requirements on “busy” custody officers across the three custody suites originally selected for the trial. Consent would, therefore, be sought in two stages: agreement to consider the experiment and return voluntarily to the police station at an appointed time (but with the clear warning that failure to attend would trigger prosecution); fully informed consent, following a detailed, structured interview with an offender manager.

Given that the operational considerations were accorded priority in this aspect of the design, the approach posed an obvious and known threat to treatment

integrity. Designs in which consent was obtained after randomisation – as in the Bethlehem Restorative Justice RCTs (McCold and Wachtel, 1988) – had produced very low levels of integrity. The research team debated the extent of the risk and the level of acceptable refusals (DOC/TP/4), given that all those assigned to the Turning Point treatment were to be analysed on the basis of their intended treatment, whether they declined treatment, failed to attend the initial appointment or declined to sign a contract (Hollis and Campbell, 1999). A key factor in the debate was the author's prior experience as a custody officer and recognition that custody officers would be reluctant to go through a full explanation of the trial in a custody environment. The research team agreed, instead, to monitor the refusal rate through the preparatory stages and return to the issue in the event that it exceeded 15% (DOC/TP/4). As the CONSORT diagrams show, the refusal rate never exceeded 10%. This presented a marked contrast with the high levels of refusals in McCold and Wachtel's studies. The difference can probably be explained by the fact that the Turning Point offenders had a considerable incentive: the offer of avoiding a criminal conviction and the subsequent criminal penalty in the event of successful compliance. There was also a relatively short timescale – 24 to 48 hours at most – between the initial pre-consent and full consent in a meeting with the offender manager. In contrast, the process of arranging a restorative conference usually creates a considerably longer "cooling off" period.

4.6.2. (f) Victims

In addition to the process for obtaining consent from the offender, the project board also needed to consider the extent to which victims were to be consulted. The Code of Practice for Victims of Crime, which was introduced in 2006, provided a statutory framework for the obligations to victims in the criminal justice system. The Code placed an obligation on the police and CPS to ensure that vulnerable victims or victims with special needs were identified and their needs taken into account and that all victims were informed about significant events or changes in the investigation. The latter specifically included the decision to caution or divert an offender or to charge an offender to go before the court. However, the decision to divert remained with the

police and CPS, having taken account of the victim's wishes. Within the experimental design, the Damocles Project Board recognised a clear difference between crimes with no personal victim, property crimes and crimes against the person. They concluded that there should be a sliding scale of obligation on the officers dealing with the crimes, with a greater obligation on the police to consult the victim in crimes against the person prior to the final decision and agreement on the conditions.

However, the victim experience of the experiment remained a live issue throughout the trial, resurfacing in the debrief of the preparatory stages and prompting an exploratory survey of victims, which, in turn, led to a parallel experiment testing a radically revised approach to engaging and consulting victims (Slothower, 2014). As the extent of victim dissatisfaction with some of the early referrals to Turning Point became apparent, the treatment of victims and compliance with the Code emerged as the single most important risk to the trial (DOC/TP/6 and DOC/TP/7). It prompted a series of measures ranging from training and briefing to the creation of a dedicated team and victim script (DOC/TP/8 and Slothower, 2014).

4.6.3. (g) Summary

Overall, the legal and ethical considerations were complex and interconnected. Miles and Huberman (1994) suggest that research tends to involve dilemmas and conflicts and that negotiated trade-offs need to be made as the research progresses. Given that, as with the issue of offender consent prior to randomisation, such trade-offs can create predictable risks to the integrity of the trial, it is apparent from Operation Turning Point that such assumptions should be properly tested before the trial goes into full operation.

4.6.4 Piloting and Pre-testing:

Sherman (2010) strongly recommended a careful, staged process of preparation and pre-testing. Lessons from Dunford (1990), Earle (1973) and Dennis (1988) suggested that this pre-test phase should include a dry run of

the eligibility screening, random assignment processes and eligibility overrides. The Damocles project board agreed to two stages of pre-testing, supported by training of as many of the key staff in custody and offender management as possible (DOC/TP/10). Stage 1 was intended to test the Cambridge Gateway. In order to do this the Gateway was to be set to assign all eligible cases into the control condition. A month before Stage 1, the key staff were bused to Cambridge and provided with an all day workshop centred on the research proposal and case studies to allow a simulated walk through of the procedures in custody and offender management.

4.6.4 (a) Pre-test decisions for the Project Board

Before the training and Stage 1 could be commenced, the project board had to settle two key operational issues: firstly, the random assignment process and delivery; secondly, the treatment delivery. The debates on the project board and between the project team and research team centred on the continuum between maintaining normal operations framed by a clear strategic direction from the Chief Constable in favour of enhancing operational officers' professional discretion and the requirements of internal validity in the experiment. It was apparent from pre-test calculations of predicted case flow (DOC/TP/9) that restricting the experiment to first offenders arrested within two Birmingham Local Policing Areas (LPA) might require running the experiment for more than 20 months in order to achieve the 400 cases which effect size calculations suggested would be required (DOC/TP/11). Experience (Strang, 2012) suggested that case flow predictions were likely to overestimate the real flow, which meant that the project board also needed to consider options to extend the trial to other LPAs as well as extending the eligibility criteria beyond first offenders. Extending the duration of an experiment or expanding its scale or sample size both presented well recognised risks to treatment integrity: the former from creeping cross contamination (Clarke and Cornish, 1972); the latter from greater heterogeneity in the sample and treatment delivery (Weisburd et al., 1993)

4.6.4 (b) Random Assignment process

Even restricting the experiment to Birmingham South and Birmingham East meant that the experiment would be running in three custody centres – Bourneville, Kings Heath and Stetchford – with at least 30 custody officers involved. In the process that the project board finally agreed (DOC/TP/10), those 30 officers would be solely responsible for opening the Gateway and completing the eligibility screening each time that they determined that a case met the criteria for prosecution. Randomisation would, therefore, be managed without the presence or direct involvement of a researcher. However, the Gateway design was adapted to provide an email of of every completed Gateway entry to designated members of the project and research team. The intention was that this would enable both tracking and follow up to maintain high levels of integrity without requiring close monitoring in the custody suites. This came closest to Braucht and Reichardt's (1993) vision of a computer based model of random assignment. However, as the scope and geography was expanded in Stage 3 to include Birmingham West and Central and Birmingham North by September 2012, the total number of custody staff engaged in randomisation rose to 90 (Hobday, 2015), an unprecedented number of decision-makers to be entrusted with the management of the random assignment process in a police RCT.

4.6.4. (c) The Treatment delivery model

There was a similar dilemma in the design of the treatment delivery process. Sherman (2010) and Sherman et al. (2015) drew on the contrasting experiences in the RISE and UK RJ experiments to recommend a small, dedicated, treatment delivery team. However, in the context of severe cuts to operational budgets and a research fund which was limited to the research costs, the project board decided that a dedicated team was unachievable. They decided, instead, to give the task of assessing, agreeing, supervising and, where necessary, enforcing the treatment plans to police officers within the local area Offender Management Teams (OMT). These officers were tasked with the management of medium and high risk offenders as part of the

Integrated Offender Management partnership with the local Probation and other agencies. The project board preferred the option of using the OMT officers because it was felt that they had experience of managing offenders and should have the requisite skills and training for the role envisaged in the trial. They rejected the alternative of relying on local response team officers because, whilst this would have enhanced the number of officers available to manage the workload, it would "potentially increase the inconsistency of response" (DOC/TP/10:3). There was insufficient attention devoted, at this stage, to the potential differences in service provision between adults and juveniles. As the trial started, it became necessary to address this and, as something of an afterthought, the juvenile Turning Point process was devolved to the Youth Offending Service.

4.6.4. (d) Pre-Test Training

The pre-test training session at Cambridge, which drew together Custody Officers, OMT and the project and research team concentrated on consistency of response, particularly to screening and treatment options. Two key issues were carefully rehearsed: the upper threshold for eligibility which required that custody officers could assess whether the offender's actions were likely to result in a custodial sentence; the potential treatment options. The former required the custody officers to be able to access and interpret the sentencing guidelines. The latter, whilst informed by the outline strategy in the research proposal, needed developing with clear parameters as to dosage, delivery and acceptable proportionality when compared to the likely sentence a court might award an offender randomised for the same offence into the control condition.

It was in the debates around these options at the Cambridge training that the custody officers and OMT members decided that the experiment should be called "Operation Turning Point", because it appeared to characterise the shift away from enforcement towards pre-court rehabilitation and support for desistance better than the original "Damocles" title. One officer summed up the discussion in an email to the research team after the training:

> "As a policeman of some 11 years it is easy to become entrenched with the view that anything other than hanging as a punishment is a 'softly, do-gooder' approach, however on returning home I have given the experiment and the facts that you provided us with some thought and you have made me really question my own long-held beliefs" (DOC/TP/14).

In the subsequent interviews, one of the custody staff reflected on the impact of the initial training:

> "But we are all invited down to Cambridge for the day and we were all treated very well, and we were given a fantastic day when [the research team] give us all the input and we felt as though because that couldn't have come cheap in both time and resources, as well as costs, that it certainly made me feel that WMP and Cambridge are serious about this otherwise they wouldn't be putting all this effort in." (DOC/TP/IV/C/1)

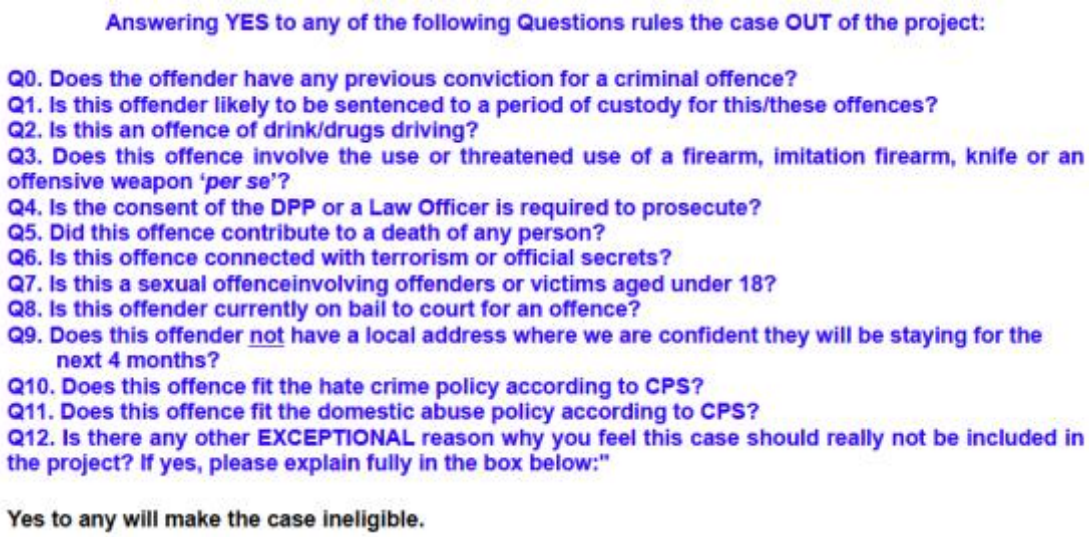
4.7 Screening for Eligibility and Random Assignment:

Ariel et al. (2012) argued that random assignment could be achieved "without tears" by using the "Cambridge Randomizer" which "is a user-friendly, safe, and cheap platform, which enables treatment providers to conduct the allocation themselves. The integrity of the random allocation procedure can be preserved, as the researcher team maintains full control over the process at the backend of the process." (p. 202). The Cambridge Randomizer or "Gateway" [5] was a critical part of the Operation Turning Point design, but was used in a very different way to the original version deployed in the first West Midlands RCT. In that trial the Gateway had essentially provided a registration system and random assignment, which was managed by a small group of

[5] It was renamed the "Cambridge Gateway", because the force Communications staff were worried that "randomiser" might create the wrong impression about the trial

trained Sergeants overseen by the researchers (Ariel, et al. 2012). For Operation Turning Point, a much more ambitious eligibility screening and data gathering process was added to case registration and and random assignment.

The initial, Mark 1.0, Gateway (Figure 4.5) was drafted by the research and project team, shared with the project board and approved by the Crown Prosecution Service (CPS) member of the project board. Prosecutor engagement in its design and approval was seen as important by the project board in order to provide legitimacy and remove potential obstacles to the experiment. The focus of Mark 1.0, which was used in Stage 1, was on ensuring that ineligible cases were excluded. Custody Officers were instructed that the Gateway should be opened and all the questions completed every time a case met the threshold for prosecution. Compliance with this requirement was seen by the research team as vital to ensure that cases were consistently screened. The two Custody Inspectors responsible for supervision of the custody staff were tasked with tracking Gateway records against the custody database.



Answering YES to any of the following Questions rules the case OUT of the project:

Q0. Does the offender have any previous conviction for a criminal offence?
Q1. Is this offender likely to be sentenced to a period of custody for this/these offences?
Q2. Is this an offence of drink/drugs driving?
Q3. Does this offence involve the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon '*per se*'?
Q4. Is the consent of the DPP or a Law Officer is required to prosecute?
Q5. Did this offence contribute to a death of any person?
Q6. Is this offence connected with terrorism or official secrets?
Q7. Is this a sexual offenceinvolving offenders or victims aged under 18?
Q8. Is this offender currently on bail to court for an offence?
Q9. Does this offender not have a local address where we are confident they will be staying for the next 4 months?
Q10. Does this offence fit the hate crime policy according to CPS?
Q11. Does this offence fit the domestic abuse policy according to CPS?
Q12. Is there any other EXCEPTIONAL reason why you feel this case should really not be included in the project? If yes, please explain fully in the box below:"

Yes to any will make the case ineligible.

Figure 4.5: Turning Point Gateway Mark 1.0 October 2011 (DOC/TP/5).

The custody Inspectors were also asked to focus on the custody officers' use of the final, "exceptional", exclusion question (Q12). The Gateway required them to set out their reasons for exclusion, allowing the Inspectors to

scrutinise and hold custody officers to account. Dennis (1988) had emphasised the importance of providing “overt” overrides to reduce the potential for “covert” overrides, which he argued were likely to be much more problematic for bias. Tracking the Gateway exclusion question (Q12 in Version 1.0, Q14 in Version 2.0 and Q15 in Version 3.0) became a key part of the programme and research team’s approach in all four phases of the experiment (Hobday, 2015).

The initial version of the Gateway and the consistency of the screening judgments were reviewed after Stage 1, in the early stages of Stage 2. The research team set aside a full day for a structured debrief (DOC/TP/11), a key part of which involved engaging a group of Custody Officers in discussing and debugging the operation of the Gateway. A number of key points emerged: the wording of the original Gateway was seen as ambiguous and not written in “custody language”; the Custody Officers wanted offenders who would be likely to receive driving penalty points to be screened out; the Gateway was not yet stable in operation. The Custody officers suggested that there was scope to include offenders with a single previous conviction. The process and the subsequent design changes appeared to have had a significant impact on the custody staff:

> “the Cambridge research team will change things if one of my teams or one of the OM comes up with a suggestion to say I don’t think that really works that way of doing it can we do it this way? They will change it. The guys have seen that happen and you get a massive buy in from them, as they feel much more participative of the whole thing.” (DOC/TP/IV/C/2)

The research and project team were also able to draw on the CPS assessment of the consistency of screening decisions in the first two months of operation (DOC/TP/12). The CPS looked at four areas: whether the case met the Turning Point criteria; whether CPS would have been likely to prosecute the cases assigned to Turning Point in the event of a breach; whether the conditions imposed in the treatments in Phase 2 were

proportionate; the appropriateness of specific conditions. A summary of the CPS responses are set out in Table 4.2. The CPS judged that in over 60% of cases the Custody Officers were clearly compliant with the screening, but that a small number of more serious cases and cases where an out of court disposal could have been considered had been included. The latter were all juvenile cases, where, as a result, the CPS suggested that they would also have been reluctant to prosecute. However, the CPS judged that the conditions proposed for the offenders in Turning Point were generally appropriate (88%). The CPS, like the Custody Officers, suggested that the project board should consider including offenders with a single previous conviction.

	Compliant with TPP screening criteria?	CPS likely to prosecute if breached?	Appropriate conditions?
Yes	14	18	22
No	9	7	3
Unclear	2	0	0
Comments	Cases that were outside criteria were either judged likely to result in a prison sentence or to be more suitable for Out of Court Disposal	The cases that CPS judged “unlikely” were primarily juvenile cases where a reprimand/final warning was judged more suitable	CPS commented on the absence of Restorative Justice conditions

Table 4.2: CPS assessment of the consistency of decision-making in Phases 1 and 2 (DOC/TP/12).

With the benefit of the debriefing session, the CPS review, the ongoing tracking of Gateway compliance and Q12 exclusions, the research team redrafted the Gateway to produce the Gateway Mark 2.0 (Figure 4.6). The revised version incorporated comments from the Custody Officers. Cases involving driving penalty points that could lead to disqualification were excluded. The project board also expanded the eligibility criteria to include offenders with a single conviction, provided that five years had passed without offending (or two years if they were a juvenile). This decision was driven primarily by the need to increase the case flow, which had been significantly slower than anticipated at that stage.

Ramdomizer Case Number:

Custody No: Custody Officers
Collar No:
1. Is this a case involving a single defendant? Yes/No
2. If no, please list the custody record numbers for all the defendants
Defendant 1:
Defendant 2:
Defendant 3:
Defendant 4:
Defendant 5:
3. Do any of the defendants have any previous convictions for a criminal offence? Yes/No
4. Where the defendant(s) have only one conviction, is it more than more 5 years (adult) or more than 2 years (Juvenile) since that conviction? Yes/No
5. Is the defendant (are any of the defendants in a multiple case) under 18's ? Yes/No
6. Is this offender (are any of the offenders, where there are multiple defendants) likely to be sentenced to a period of custody for this/these offences? Yes/No
7. Is this a driving offence, including drink/drugs driving, that is likely to lead to a driving disqualification? Yes/No
8. Does this offence involve the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon 'per se' ? Yes/No
9. Is the consent of the DPP or a Law Officer is required to prosecute? Yes/No
10. Did this offence contribute to a death of any person? Yes/No
11. Is this offence connected with terrorism or official secrets? Yes/No
12. Is this a sexual offence involving offenders or victims aged under 18? Yes/No
13. Is this offender (are any of the offenders, if multiple defendants) currently on police bail, bail to court for an offence, on licence or serving a court-imposed sentence in the community? Yes/No
14. Does this offence fit the hate crime policy according to CPS? Yes/No
15. Does this offence fit the domestic abuse policy according to CPS? Yes/No
16. Are there any other reasons to exclude the case? Yes/No
If Yes, give reasons:

If answer to all (apart from Q1) is No,
then ELIGIBLE

Figure 4.6: Operation Turning Point Gateway Mark 2.0 February 2012-April 2013 (DOC/TP/13)

The project board also agreed (DOC/TP/16) that the experiment needed to be extended to the other two Birmingham LPAs – Birmingham West and Central and Birmingham North – in a phased approach, which would be preceded by a second wave of training of the custody staff and offender managers in the original LPAs. This second training input was scheduled to be complete in

May 2012 before the experiment moved to go into full live operation on June 1st 2012. The programme for the second phase of training (DOC/TP/13) involved a reinforcement of the theory and rationale for the experiment together with an interactive session in which real case studies from Stages 1 and 2 were to be used to encourage debate and draw out issues that had emerged.

4.8 The Treatment Procedure

By the time the second phase of training took place, Stage 2 of the experiment, in which the Gateway assigned all eligible cases to Turning Point, had been running for more than four months. More than 40 cases had been put into the treatment condition. This meant that not only could the CPS review the appropriateness of the early condition setting, but the newly arrived field researcher[6], was able to analyse the treatment records and conduct an exploratory survey of the victim and offenders' experience of Turning Point (DOC/TP/17&18).

The treatment procedure and the protocol underpinning it had been agreed by the project board before the start of Stage 2 and set out in the Crimport (Appendix 4), which had been published on the Cambridge Institute website[7] after approval by the board. The approach had been designed to meet the hypothesis set out by Sherman and Neyroud (2012) that celerity and certainty supported by a range of desistance based, rehabilitative conditions were likely to be more effective in reducing reoffending by low harm offenders than the control condition of court prosecution. Having decided that a special treatment team was not a viable option, the project board placed the responsibility for assessing adult offenders and agreeing and managing the treatment plans on the two Local Policing Area Offender Manager Teams (OMTs). Young offenders would remain with the Youth Offending Service (YOS), a partnership with Birmingham City Council. Each team of OMTs was

[6] Molly Slothower, a PhD student the University of Maryland.
[7] http://www.crim.cam.ac.uk/research/experiments/rex-post/operation_turning_point.pdf

supervised by a Sergeant, who was expected to oversee the treatment plans and compliance with the experiment. The YOS had their own local manager. The OMT's primary role was to work within the framework of "Integrated Offender Management" which was designed to provide "a cross-agency response to the crime and reoffending threats faced by local communities by managing the most persistent and problematic offenders identified jointly by partner agencies working together" (Home Office, 2015:1). The YOS worked within a wider youth justice framework (Ministry of Justice and Youth Justice Board, 2013), in which a statutory partnership under the leadership of the local authority was responsible for the coordination and delivery of local youth justice services, including managing and supervising young offenders referred to them.

There had been debate in the project board at the outset at the extent to which the OMTs were the appropriate team and their operational resilience to be able to absorb the Turning Point workload alongside their normal work (DOC/TP/10). The estimated and emerging case flows suggested that the workload was sustainable. The YOS was not engaged in the original discussions about the experiment and, although they agreed to participate and became full members of the project board, there were significant differences between the juvenile and adult treatments, which were only partly explained by the different practice and statutory frameworks. In particular, the YOS model of working placed a much greater emphasis on regular face-to-face meetings than the OMTs (DOC/TP/23). The Gateway model adopted for Stage 4 duly recognised the importance of the different approaches and introduced block randomisation of adults and juveniles to enable separate evaluation.

However, whilst there were some differences in the treatment approach, the treatment procedures were intended to be similar. Offenders who had been screened by the Gateway and were found to eligible were assigned by the Gateway algorithm to either control (prosecution) or the Turning Point treatment. Those assigned to control were charged as normal. They were "blinded" as they not told that they were part of a trial. On the other hand,

those assigned to Turning Point were informed that they were being offered the opportunity to be part of a trial. Subject to their initial consent, they were asked to attend an appointment at the police station with a member of the OMT. The original intention, driven by “celerity”, was that this would be within 24 hours, seven days a week. However, the operational realities of providing such cover within the relatively small team of OMTs meant that, by Stage 3, it was accepted that 48 hours was more attainable. From time to time, communication difficulties between custody and the OMTs and YOS contributed to offenders (6 in Stage 3) failing to attend their initial appointment. It is notable that such logistical difficulties were also highlighted in both the RAND (Disley et al., 2016) and MoJ studies (Ames et al., forthcoming).

When the offender turned up for that initial appointment, they were to be met by a police officer from the OMT (if an adult) or a YOS member (if a juvenile). The meeting was intended to allow the OMT or YOS member to discuss the offence and the reasons for offending with the offender, to explain the Turning Point experiment and encourage the offender to suggest and agree a set of conditions, which could reduce the likelihood of them offending. A standard Turning Point pro-forma was developed and tested during Stage 2, which was designed to enable the process to be recorded and monitored. By the end of the interview, the offender and the OMT and YOT team member were required to agree a “Turning Point Contract”, which would last for around 4 months and would always include two generic conditions – to comply with the contract and not to re-offend – together with two or more conditions that had been agreed with the offender and which were intended to be tailored either to their offending pathways or the specific risks associated with the offence they had committed.

Once the offender had signed and agreed their Turning Point contract, they had to comply with it or risk being prosecuted for their original offence. Tracking compliance was the responsibility of the OMTs and YOS. Turning Point Offenders were flagged on the local West Midlands ISIS Custody system, the West Midlands CORVUS Offender manager system and on the

Police National Computer, so that if they were re-arrested they could be breached where an offence was subsequently charged. A major concern in other cautioning trials, particularly the trials of restorative justice (Strang et al., 2013), had been the challenge of ensuring that the conditions are actually enforced. In this experiment, the onus was placed on the offender producing pre-agreed evidence of their compliance, in order to enable the OMT or YOS supervisor to approve the completion of their contract.

With the exception of the small group of offenders who refused (Figures 4.1 and 4.2), in every case in Stages 2,3 and 4, where offenders were assigned to Turning Point, a contract was agreed. As such, this technically complied with the Crimport specification of the intended treatments, which was set out as follows:

> "8.1.1.1 All the subjects allocated to treatment must have a "diagnosis meeting" with a member of the offender management team within 48 hours of arrest and must sign a "turning point contract" setting out the actions, including no reoffending, which they have agreed to following on from the "diagnosis meeting". Cases where these two conditions are not applied cannot be considered to have met the conditions of the treatment." (Appendix 4)

Therefore, the fact that an offender had agreed the contract, including a requirement not to re-offend, following a meeting with a member of the OMT, and was "at risk" of prosecution for breach was sufficient to meet the strict interpretation of the Crimport requirements. That interpretation covered primarily the "Sword of Damocles", deterrence based approach, which had been referenced by Sherman and Neyroud (2012) first and foremost to Dunford's (1990) Omaha RCT, in which the findings had supported the effectiveness of holding the imminent threat of punishment over the offender. However, as the CPS review of the early conditions showed (Figure 4.2), those core requirements were not operationally sufficient. The conditions needed, in addition, to be specific enough to enable officers to determine whether they had been breached and present the evidence in a way that was

capable of securing a conviction in court.

As the revised title of the experiment – “Operation Turning Point” - indicated, there was a wider interpretation of the treatment requirements which included the intention to adopt a more rehabilitative, restorative and reparative approach designed to support desistance. The CPS had offered an assessment in their commentary on whether the early conditions were “appropriate” to the offender and their offending. As the experiment developed an increasing emphasis was placed on moving beyond the narrow interpretation of the treatment requirements. Partly this was driven by the police members of the project board and the offender managers who expressed a growing interest in understanding which conditions worked best with which offenders. Partly, it was driven by the police offender managers and their supervisors, who started asking for clearer advice about “evidence-based conditions’ (DOC/TP/17). It was also supported by the emerging evidence from the survey work with victims, which identified that the credibility of police efforts to prevent reoffending was a key factor in underpinning their confidence levels in the police handling of the case (Slothower, 2014a).

However, the most powerful driver of change was the feedback from the field researcher about the quality of the conditions in the initial plans (DOC/TP/17). This analysis was used repeatedly in the second round of training, in briefings to the project board and the Chief Officer team (DOC/TP/20) and in the one to one and small group meetings. The project team – both West Midlands Police and Cambridge research team – revised the framing of the treatment procedure for Stage 3 from the CRIMPORT model to a “SMART” model supported by a paper provided by the research team on the evidence for the emerging set of conditions (Neyroud and Slothower, 2012: Appendix 4). The extent of this shift was embodied in the change from the phrase “Endless possible tactics” (DOC/TP/21) used in the earlier training presentations towards an emphasis on high levels of compliance with “SMART” evidence based conditions in the presentations for the second and subsequent phases of training. This required conditions set by the OMT and agreed by offenders to be “Specific, Measurable, Attainable, Relevant and Time-bound” (Doran,

1981) and drawn, as far as possible from the menu of evidenced tactics (Neyroud and Slothower, 2012: Appendix 4).

As such, it was not only intended that it should be clear to the offender what was required of them, but also so that there was a fair and transparent means of assessing a breach against a set of conditions which were, as far as the literature allowed, supported by evidence of their potential effectiveness. This was seen as being important for a number of reasons: the need to evidence a breach to the CPS and courts; the need to demonstrate to victims that a credible effort was being made to prevent reoffending; as a result of both, a more legitimate approach, referencing the body of research indicating the importance of procedural justice in criminal justice processes (Tyler, 2007).

The field researcher on the research team examined the results of applying the new treatment approach and the effectiveness of the compliance strategies (Slothower, 2014b). The first part of this study involved the field researcher, author and a member of the project team coding and inter-rating 202 Turning Point contracts from Stages 1-3 against the SMART criteria. Slothower (2014b) divided the results into three phases:

- **Discretion only or "Leave it to the officers"**: In this, the first phase, which prevailed during Stage 2 of the experiment, Slothower found that only around 50% of the conditions were specific and less than a third had set out a measurable dosage of treatment.
- **Training**: after the officers received the second phase of training prior to the commencement of Stage 3, they were encouraged to focus on SMART conditions by being provided with examples of good and bad conditions from Stage 2. Their supervisors were also encouraged to check and feedback on conditions to their OMT officers. This raised the levels of compliance to more than 70% and just under 60%. A substantial improvement, but one which, nevertheless left more than a third of contract conditions short of the desired standard.

- **Training and provision of recommended conditions**: The addition of guidance to supplement the training appeared to produce little if any improvement. This finding was consistent with Fox (2012) whose analysis of police officers' response to written policy guidance suggested that, whilst they were an inevitable requirement of operational bureaucracy, they were not particularly effective at achieving changes in practice.

Thus, by the middle of Stage 3, with two additional Local Police Areas coming into the experiment in July (Birmingham North) and September (Birmingham West and Central), the treatment delivery, supported by training, guidance and supervisory feedback was achieving apparently high levels of compliance with the original CRIMPORT standard and increased levels of SMART compliance. However, the latter was still – at 78% specific, 87% attainable and 61% measurable – on average below the 90% threshold we estimated as "Good" in Chapter 2.

4.9 Stage 3: Uncovering problems with Treatment Integrity:

Even so, at this stage in early 2013, the experiment appeared to be largely "on track". The case flow had significantly increased with four LPAs live. The compliance with the treatment procedures had been improved and randomisation appeared to be going well. However, there were some indicators of problems emerging. In August 2012, the project manager, Inspector Jamie Hobday and the researcher conducted an audit of custody officer's compliance with the requirement to screen all offenders whom they had decided to charge. They found that although "custody officers had informed us from their deliberations we were at about 90%" that, in fact "we were actually at 50% in July and have now dropped to 35% for the first week in August" (DOC/TP/22). Two reasons were put forward for this: firstly, the project manager concluded that custody officers were "pre-screening" cases, either because they did not want them to enter the experiment or because they had concluded that they would be excluded anyway; there were some

reported technical problems with the Gateway. However, the project board, in discussing the issues, concluded that a third reason – disruption to the custody system through the Summer period with temporary Custody officers being bussed in from outside the experimental area – had been the major factor.

The project team tasked the custody Inspectors with more detailed tracking and follow up to ensure compliance. They commissioned an interim report for the Chief Officers, which was prepared by the author and the field researcher (Neyroud and Slothower, 2012). It highlighted the progress, issues to date and lessons learnt to that point. It stated that "Operation Turning Point experiment has been set up and carefully implemented in stages. The Operation is now set to achieve its target sample size by the second half of 2013" (p.42).

Through the autumn of 2012 and winter of 2013, the field researcher was not in the field, because of UK visa restrictions. However, there was a now well established system of Randomiser data, custody Inspector and OMT Sergeant oversight supported by an internal data analyst who was tasked with maintaining the case records as they came through. Fresh impetus had been put into the system of tracking after the problems identified with screening and the project board was meeting at least monthly to review progress. Despite this, just as the absence of the field researcher in one of Dunford et al.'s (1982) field sites had proved a problem, so it proved again, despite the very considerable progress that had been made in Stages 1-3.

When the field researcher returned in March 2013, there was a detailed review of progress including a forensic review of the treatment integrity of the cases processed so far. Starting with the positives, the researcher observed that two of the major threats to the project identified in the pilot phases had been "largely addressed": "people are more or less on board"; the conditions are looking quite tight" (DOC/TP/23). Having identified serious problems with attrition and some issues with treatment delivery, the researcher observed that "the Cambridge Gateway ("randomiser") for the Turning Point Project was

designed to provide guidance for a large group of people to make complex vetting decisions. However, it has never before been used by *such* a large group of people, to make *such* a complex vetting decision" (DOC/TP/23: 2). Noting that more than 90% of the charged cases were now being entered in to the Gateway, the researcher observed that, given that only around 5% of all cases being entered into custody were being screened in, this meant that complying with the screening was proving irksome in a busy custody environment. Nevertheless, despite this, Hobday's (2015) study of the use of the override questions suggested that there had been continuous improvement in the credibility of grounds for exclusion.

Whilst none of this would, therefore, suggest a serious problem, the level of attrition in the sample between treatment and control was a 'show stopper'. Several issues had contributed to produce a treatment integrity that the research team now assessed as being below 60%. The problems could be divided into two: technical faults with the Gateway; human faults with the use of the Gateway. Critical in the first category was an unintended secondary override box (Question 14.5) that appeared on screen after the randomisation decision and offered an opportunity to exclude cases on which the allocation decision was now visible to the custody officer. Furthermore, the additional question was not visible to the researchers in the Gateway emails of each decision. The research team discovered that 48 cases were wrongly excluded post randomisation by this means.

On the human side, the research team found that cases that had been allocated by the Gateway to Prosecution had nevertheless been sent to Turning Point by the Custody Officers. It was unclear whether they misunderstand the Gateway instruction that a case is "Eligible – assigned to prosecution" or simply ignored it. The latter appeared to be just as likely because it also became clear that a number of custody officers had sought to manipulate the Gateway – largely in favour of securing a Turning Point contract for the offender - by 'rolling the dice' for a second or even third time. The Gateway was not, at this stage, programmed to prevent repeated attempts with the same custody number. Although most of the problems

appeared to be in the custody suite, there were also further problems downstream with cases that had been allocated to Turning Point being returned to custody officers for either prosecution or caution contrary to the Gateway allocation decision.

The scale of the attrition, combined with the continuing, but less severe, shortfalls in treatment delivery, was summarised in emails between the research team members discussing how to manage the scale of the problem with the project team and the force. The author advised that the research team share a “short objective note…and a deep breath…. I have taken several since discovering Q14.5” (DOC/TP/24:1). The field researcher commented “our WMP counterparts have handled the setback with tremendous grace, largely because of the feeling that we are all in this together, and we are all doing everything we can to make it work” (DOC/TP/24:1). However, the “setback”, after nearly two years of work to set up, pilot, pre-test and running the experiment was considerable for both the research team and the force. The research team concluded that the data generated in Stage 3 had so high a risk of bias as to render it inappropriate to use it. The research team, therefore, recommended that the only credible solution was to restart the experiment, with a revised Gateway, a new treatment approach, a fresh round of training and feedback and enhanced data analyst support.

Amongst the analysis of Police RCTs in Chapter 2 this author could only find two experiments which appeared to have been revised and restarted after early problems (Earle, 1973 and Weisburd and Green, 1995). Of these, only Earle’s experiment involved a complete restart with lessons – particularly revised screening – incorporated. On the whole, as Weatheritt (1986) observed, police “experiments” are expected to succeed and, if they do not, are likely to disowned and abandoned. Earle’s study appears to have been able to rebound largely because Earle was very senior in the department and its head, Sheriff Pitchess, was clearly committed to reform.

There were parallels with this in Turning Point. The Chief Constable had explicitly committed to evidence-based policing as a key strategic aim. Turning Point had been very publicly presented as a key part of that approach. The Chief Officer team had also secured the support of the newly elected Police and Crime Commissioner to the project. Moreover, the Ministry of Justice was working up to a launch of its proposed reform of diversion (Ministry of Justice and College of Policing, 2013), which would draw heavily on the Turning Point design, which had been shared in detail with the steering group of senior national officials and had engaged their support. Both at national and local levels, the force was strategically and operationally committed to move forward with Turning Point and had embraced evidence based policing. The project board elected to restart rather than withdraw.

4.10 Stage 4: Restarting the Experiment

For the project team and researchers, the challenge was now to ensure that Stage 4 succeeded in producing a successful experiment with more than 400 cases randomised and treated with a high level of treatment integrity. As the Turning Point Treatment Integrity "storyboard" (Appendix 5) shows, the focus was on three key areas: revising, re-testing and restarting the Gateway; introducing a "prescribing tool" to support higher levels of compliance with SMART evidence based conditions for the contracts; retraining and rebriefing key staff in a positive and constructive way to achieve compliance. Alongside this was an urgency to secure additional data analytical support from the research team to replace the internal data analyst whose tracking and monitoring had proved insufficient to assist the early detection of problems.

The Gateway Mark 3 (Figure 4.7), which went live in April 2013, was designed to rectify the problems with Mark 2 and to address the feedback from custody officers about unnecessary workloads from filling in the whole form for every case. There were five key changes: the Gateway was set to screen as 'ineligible' as soon as any of the criteria were not met; Question 14.5 – the unintended secondary exclusion, post randomisation, was removed; the

Gateway was divided into the core screening criteria on the first page and more explanatory information on the second; once the Gateway had determined that a case was eligible and randomised it, the third screen provided a step by step checklist on what needed to happen next; the Gateway was set to block randomise cases into those juvenile and adult offender cases with and without a personal victim, in order to enable an RCT within the RCT on victim treatment and satisfaction (Slothower, 2014a). The new Gateway was also built on a commercial platform rather than in house in an effort to secure greater resilience and faster resolution of any issues identified.

PHASE I REGISTRY: Here is the data that was submitted:

Ramdomizer Case Number: **60254**

Custody No: **fa/939/13**

Custody Officers Collar No: **3681**

What kind of case is this? **Adult without personal victim**

1. Is this a case involving a single defendant? **Yes**

2. Does the defendant (any of the defendants) have any previous convictions for a criminal offence? **More than one**

3. Is this offender (are any of the offenders, where there are multiple defendants) likely to be sentenced to a period of custody for this/these offences?

4. Is this a driving offence, including drink/drugs driving, that is likely to lead to a driving disqualification?

5. Does this offence involve the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon 'per se'?

6. Is this a sexual offence?

7. Is this offender (are any of the offenders, if multiple defendants) currently on police bail, bail to court for an offence, on licence or serving a court-imposed sentence in the community?

8. Does this offence fit the hate crime policy according to CPS?

9. Does this offence fit the domestic abuse policy according to CPS?

10. Is this an immigration offence?

11. Is(are) offender(s) already on a Turning Point Plan, or have they been in the past?

12. Is this offence connected with terrorism or official secrets?

13. Is an order of the court required, and a similar outcome cannot be achieved within a Turning Point Plan (e.g. Sex Offenders Order)?

14. Is the consent of the DPP or a Law Officer is required to prosecute?

15. Did this offence contribute to a death of any person?

16. Does the offender have a confirmed address suitable for the service of summons?

Is there any other EXCEPTIONAL reason to exclude this case from Turning Point?

Figure 4.7 Operation Turning Point Gateway Mark 3: April 2013-July 2014: (DOC/TP/19)

The research and project teams then debated the strategy for explaining the new direction and for re-training and feeding back on lessons so far. Two potential approaches were debated: targeting only the “problem custody officers” or re-training all the custody staff. The latter was agreed as the better approach with a strategy set as:

> “(a) thank them for their continued participation (b) reflect some of the continuing good news stories (c) highlight some emerging problems (d) make the connections to the cautioning review and comment that our research suggests that they take the decision to divert very seriously and we hope to be able to say this to the review (e) focus on the improvements we are hoping to make to the randomiser and the help that they need to give us to finish this experiment effectively”. (DOC/TP/24:2).

Furthermore, the sessions – individual small groups one at a time - would be supported by one of the OMT supervisors in order that the whole story of the Turning Point process could be reflected back to the custody staff. The tone and tenor of the sessions was to avoid making a drama out of the crisis, emphasise the progress and lessons learnt so far and the fact that many of the changes in Stage 4 arose from feedback from the practitioners.

The project and research team also invested in more feedback from newsletters (DOC/TP/7), updating emails and personal briefing, concentrating on stories about offenders and their response to contracts. Turning Point also became a central feature of a succession of events organised by the West Midlands branch of the Society for Evidence Based Policing, to which staff participating in the experiment were personally invited. The aim of these sessions, apart from broadcasting the project and emerging lessons beyond the experimental teams, was twofold: to encourage those teams to see themselves as innovators, blazing a trail which others would follow; to offer an opportunity to enhance their professional knowledge about and confidence in evidence based policing (DOC/TP/20).

A major feature developed at the SEBP sessions was the progress being made in the system to deliver the treatment – a deferred prosecution with conditions – which had not only proved, in the conditional cautioning model, difficult to deliver but was now of pressing interest outside the experimental area. Not only were the Ministry of Justice debating changes to OOCDs, but the Home Office and Director of Public Prosecutions had changed the rules to make conditional cautions much more accessible to the police by removing the requirement to secure prosecutorial approval before proceeding.

As we have seen, despite a significant investment in training, guidance, supervision and tracking, it had proved impossible to get above the 90% compliance threshold with SMART evidence based conditions. The OMT supervisors had formed their own peer group meetings to share experiences and discuss solutions. In discussion with the research team, they suggested moving to a “prescribing tool”, drawing on lessons from implementation in medical trials (Grol and Grimshaw, 2003). A Cochrane systematic review (Kawamoto et al., 2005) had demonstrated that “Clinical Decision Support Systems” (CDSS) were likely to significantly improve performance and consistency of practice. The review also identified that a CDSS had to be available and targeted at the right point in clinical decision-making and be supported by recording and feeding back on override decisions.

The prescribing tool that emerged as a result (Figure 4.8) provided an online replacement for the paper Turning Point contracts as well as a tool to support the OMT formulate and track the contracts created. It also provided a more available format, similar to the Gateway emails, to allow the research and project teams to monitor and track cases. With a new data analyst, Eleanor Neyroud, now tasked to create and maintain a Master Turning Point spreadsheet bringing together the data produced from the Gateway, Custody system, West Midlands Case tracking and, now, the Turning Point prescribing tool, there was an opportunity for a much tighter and more immediate capacity to track in real time.

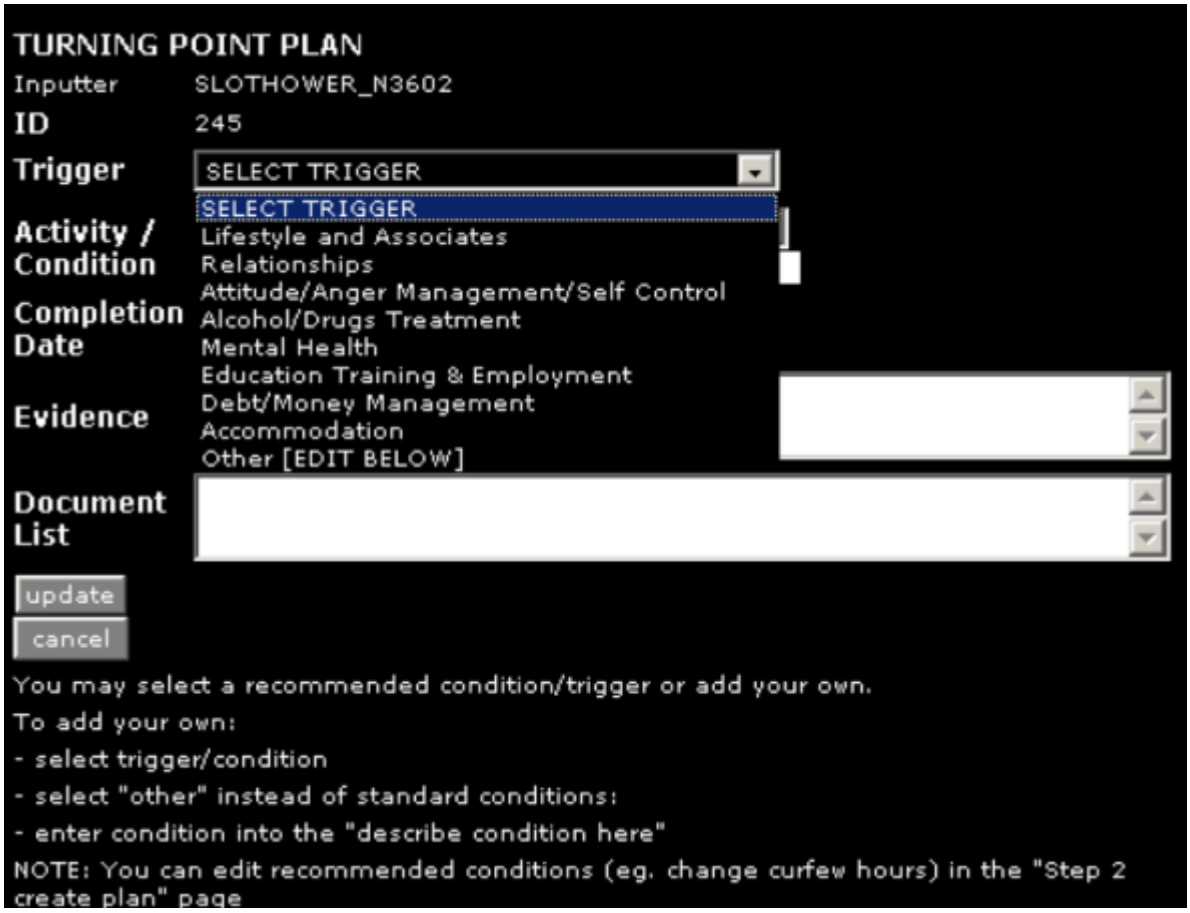


Figure 4.8: Operation Turning Point "Prescribing tool"

In a further refinement, also suggested by the OMT supervisors, a CDSS tool for risk and needs assessment of offenders was also added. The tool chosen, because of the substantial research supporting its potential effectiveness, was the LS-CMI tool developed by Multi-Systems Inc. from the work of Andrews and Bonta (2010). The tool filled a gap in police knowledge and skills to undertake risk and needs assessments of offenders that was subsequently identified in a major Joint Inspectorate review of Integrated Offender Management (CJJI, 2014).

The field researcher examined the effectiveness of these two tools in a continuation of the study of the treatment delivery models in Stages 1-3 (Slothower, 2014b). The findings suggested that both had a significant potential contribution to make, but that the combination of the two produced a very high level of consistent, compliant practice. The use of the prescribing tool on its own raised the level of compliance above 90%. With the tools

combined this was very close to 100%. The prescribing tool, in particular was seen as a very important step forward, as one OMT member commented:

> "What we needed was what we have got now…which is the portal – the portal is brilliant, because everything goes in there…" (DOC/TP/IV/OMT/1)

There were still some niggling problems because like the original Gateway Mark 1 and 2, the Portal was, initially, at least, a "Beta" version, but even so it was perceived as a strong signal of progress and continued investment in the experiment at a critical time. Moreover, it was seen by another OMT member that:

> "here we have a TP portal, and our sergeant could check that, and it seemed as soon as he could check the plans the quality of the plans improved." (DOC/TP/IV/OMT/2)

Just as the revised Gateway enabled better oversight of the randomisation decisions, the new prescribing tool portal appeared to have improved supervision of the Turning Point contracts and the ongoing management of the contracts.

Indeed, after only a month of operating the new Stage 4 model, the research team noted significant progress (DOC/TP/25). More eligible cases had been entered into the Gateway. Officers in the training sessions were reporting the enhanced ease of use of the new Gateway and confidence in using it. In contrast to the substantial monthly attrition rate in Stage 3, there had been only two cases that appeared to have been mis-assigned or re-assigned. In one of those cases the custody officer had immediately phoned the field researcher to report the error – the case should have been cautioned rather than being screened. Furthermore, the issue had been quickly shared with all the custody officers in order to avoid repetition. This set a pattern for Stage 4 in which issues were more rapidly referred to the field researcher, who was

able to be in the field throughout the Stage 4 and, where necessary escalate to the project and research manager for resolution and feeding back.

Stage 4 lasted from April 2013 until July 2014, by which the required target of 400 cases had been met and surpassed. As Table 4.1 and Figure 4.2 demonstrate, the treatment integrity across Stage 4 exceeded 90%. Both attrition and treatment delivery exceeded that threshold across the 15 months. As such, Stage 4 of Operation Turning Point appears to provide a significant outlier of good treatment integrity when compared to the juvenile justice RCTs set out in Table 3.2. above. It also provides a strong contrast to the treatment integrity in Stage 3, which was more typical of the other juvenile justice RCTs. The journey from low to high treatment integrity is represented in the steps set out in the Treatment Integrity “Storyboard” set out in Appendix 5. In the next section, we turn to a discussion of the extent to which, drawing on the Storyboard and case study set out above, we can determine key lessons for achieving successful police RCTs with high levels of treatment integrity.

4.11 Discussion of the findings

At the conclusion of the analysis of police RCT case studies and juvenile justice RCTs which began this chapter, we distilled a set of the emerging key issues in Table 4.1 above. They were divided into issues concerned primarily with attrition and those more focused on treatment delivery. Drawing on the Operation Turning case study and the Treatment Integrity Storyboard (Appendix 5), Table 4.3 sets out a summary of how those issues were addressed in the Turning Point police RCT.

Attrition	Performance
• *Pre-Test: resolving legal and ethical issues - Dunford (1990)* Turning Point used a national advisory group and the local Crown Prosecution Service to provide legal advice on the design and approach. The ethical issues were managed by the Project Board. Role of Victims emerged as a key issue for pre-court decision making • *Pre-Test: dry run of random assignment process – Dunford (1990)* The Gateway was pre-tested in Stage 1 with all cases set to the control condition and then redrafted before Stage 3. • *Design and testing of the eligibility screening process – Earle (1973) and Dennis (1988)*	• *Pre-test: dry run of the treatment approach – Sherman (2010)* The treatment delivery approach was pre-tested in Stage 2 in which the Gateway was set to allocate all cases to the "treatment" intervention. The dry run was researched in detail by the field officer including interviewing the initial sample of offenders and victims and feeding back the results in the pre-Stage 3 training sessions. • *Treatment Protocol – Gottfredson et al. (2015) and Design of the treatment.* The treatment was originally set out in the CRIMPORT that was published on the www.crim.cam.ac.uk website before go-live and then expanded as the experiment progressed. The final version of the treatment was

The screening process was designed in to the Gateway and tested in Stages 1 and 2 with external review by the CPS during Stage 2. • *Random Assignment method – Braucht and Reichardt (1993), Sherman (2010) and Ariel et al. (2012)* The Random assignment was carried out by the Cambridge Gateway – a combined randomiser and screening tool. The Gateway was revised twice and in the final version – Mark 3 – produced for Stage 4 it also incorporated block randomisation. Using the Gateway was the responsibility of the Custody Officer once a case had been determined as a "prosecution". • *Designing overt overrides and securing ownership – Dennis (1988)* The Gateway was designed with a built in "overt" override which offered the custody officer the opportunity to screen out cases which they judged as having	embedded into the "portal" prescribing tool for the OMTs. • *The "treatment team" – dedicated or within "normal" operational roles – Sherman (2010)* The "treatment team" was determined to be the Offender Management Teams from 4 Local Policing Areas and the YOS (for juveniles) rather than a dedicated team. The four OMT's operated a slightly different approach: at one end one OMT relied largely on a single officer to carry out the TP work; at the other another OMT used the available OMT officers on a rotating basis. The YOS teams also rotated involvement amongst the members of the YOS dependent on duty availability. • *Training and briefing (Ames et al, forthcoming)* There was a pre-test phase of training, followed by a pre-Stage 3 phase and further refresher phase at the inception of Stage 4. There was a constant cycle of briefings, email updates, newsletters and seminars to

"exceptional reasons" for exclusion. They had to provide their grounds for exclusion and these were tracked throughout the experiment and feedback and challenge was given by the Custody Inspectors. • *Blinding of practitioners to the Random assignment – Sherman (2010)* The Custody Officers had to be able to see the allocations because they had to make the administrative decisions – to charge or make an appointment – that the allocations required. • *Separation of operational staff from control of the random assignment process – Sherman (1992) and Sherman (2010)* The Custody Staff were responsible for the random assignment. They were capable of being tracked individually by their collar numbers against each Gateway decision and each override they entered. They were also monitored for their level of compliance with the	feedback, update, disseminate and encourage. The YOS were not trained in the initial sessions but were involved in the pre-Stage 3 training. • *Process for securing consent – Sherman et al. (2015)* Consent obtained after random assignment in two phases – initial agreement to attend a meeting with the OMT and then full consent to the TP Contract. There was tracking of level of refusals throughout the trial. • *Field Support – Dunford et al. (1982)* The Field researcher was brought in during Stage 2 and was then absent for the second half of Stage 3. The Field Researcher was brought back to review Stage 3 and was present through Stage 4. The research team data analyst was brought in for Stage 4. • *Tracking of treatment delivery, duration and dosage – Quay (1977), Sherman (2010) and Sherman et al. (2014)* Initially the tracking of treatment delivery was carried

opening of the Gateway, as required, for every prosecution decision and given feedback by the Custody Inspectors • *Tracking of random assignments – including the pre-assignment screening and post-assignment attrition – Sherman (1992) and Sherman (2010)* Every random assignment decision was sent by email to members of the project board and research team. The email included a screen print of all the decisions recorded, the officer's collar number and the custody numbers. By Stage 4 a further system was in place with a Master spreadsheet update weekly with the whole story of each case from custody through to treatment and compliance. • *Tracking the "hydraulic flow" through the experiment – Strang (2012)* The Project manager and Research manager produced an updated tracker of the case-flow for the monthly project board meetings.	out by monitoring the paper based TP Plans and relying on internal data support to type up the plans and the maintain the Master sheet. In Stage 4, the "portal" or "prescribing tool" allowed data from the treatment plans to be gathered automatically and a research team data analyst maintained the Master sheets. During Stage 3 and 4, the treatment conditions applied were coded by the Field Researcher to monitor the level of compliance with the SMART–evidence based approach that had emerged in Stage 2 and was developed in Stages 3&4.

Table 4.3 Treatment Integrity issues in Operation Turning Point compared to model in Table 4.1.

Whilst there are issues which run across both attrition and treatment delivery, there are some distinctive issues which suggest that it makes sense to start by dividing the discussion between the two. The key focus of this first part of the discussion will be on the differences between Stages 1-3 and Stage 4, given that the treatment integrity outcomes between the first three parts of the experiment and the last stage were so significantly different.

4.11.1 Treatment Integrity issues: Attrition:

Starting with the "Attrition" column of Table 4.3, the Turning Point case study appears to support the emphasis that Dunford (1990) placed on settling the ethical and legal issues surrounding the proposed RCT at the earliest stage. In Turning Point these issues were highlighted in the original research proposal and debated at both the project board and the pre-test training. By the time that the trial started the major issues – the legality and proportionality of the random assignment model, the balance of the risk management, the question of whether admission was a pre-requisite for eligibility – had been agreed and tested with the Crown Prosecution Service. However, whilst they had been settled in the formal policy and in the minds of the senior managers involved in the project board, the experience of the research team was that they required continual explanation to frontline staff, attention and reinforcement in the three stages of training and briefing that followed.

4.11.1(a) Staff Turnover and TTF

One major reason for the need for constant attention, which is a consistent challenge in police organisations, was the level of turnover in the staff. This led to more than 90 Custody Officers participating directly in the trial as decision-makers in the use of the Cambridge Gateway to screen eligibility and randomise. Hobday's (2015) analysis of appropriate decision-making on the

permitted override question showed that, despite the turnover, the level of inappropriate eligibility exclusions fell at each stage of the trial, lending support to Slothower et al.'s (2015) hypothesis that a model of a continuous cycle of training, tracking, feeding back (TTF) and retraining is essential to maintaining high standards of compliance. Reinforcing this, Hobday's data shows that the largest drop in the inappropriate screening coincided with transition to Stage 4 of the experiment. The rates by stage were: Stage 2 - 28.8%; Stage 3 - 21.7%; Stage 4 - 8.9% (Hobday, 2015). Hobday was also able to show in his surveys of the Turning Point custody officers that this improvement correlated, by Stage 4, with a high level of perceived understanding of the experiment, acceptance of the justification for experimentation and "comfort" with randomisation. In the absence of a matching pre-test survey we need to be cautious about interpreting this finding. However, it is given some support by the reverse findings in Famega et al. (2016) and MacQueen and Bradford (2016), both of whom found connections between a lack of ownership in the strategy and science of the experiment and the level of compliance with the letter and spirit of the experimental process.

4.11.1(b) Victims

One legal and ethical issue was not entirely settled and came close to halting the trial – the role of victims in the decision-making process. The fact that none of the previous pre-court diversion trials had raised the issue of victims as a major obstacle or ethical issue may well reflect the relatively recent rise of the victim's rights movement in criminal justice which was given impetus in the US by the President's Taskforce on Victims of Crime (1982), internationally by the UN Declaration (1985) and by the original Victim's Charter in the UK (2004). Indeed, only more recently has the Victim's Code and legislation in countries such as Canada sought to embed victim's rights in law, including their right to participate in key decisions such as out of court disposals.

Slothower's (2014a) early exploration of the victim's responses in Stage 2 suggested that around half were less than satisfied with the Turning Point treatment process. In response, the research team proposed an innovative scripted approach to explanation and engagement of the victim, which was tested in the parallel RCT on victim treatment and produced much higher levels of satisfaction. As the field researcher put it during Stage 3 "it is critical to document the victim issue and how we address it, as well as some evidence that it has, in fact, been satisfactorily addressed" (DOC/TP/6). One of the project team summarised the importance of addressing the victim's perspective "the victim side of it has definitely been a deviation from the original scope, but it was kinda a necessary deviation. I think the beginning focus was so much on offenders that there was inevitably going to be some sort of friction if we didn't redress that balance" (DOC/TP/IV/PB/1)

4.11.1(c) Subject consent

Whilst the victim's consent had not previously appeared as a major issue in police pre-court diversion RCTs, consent to treatment by offenders had led to high levels of attrition in more than one trial and most notably in the Bethlehem experiments (McCold and Wachtel, 1986). In Turning Point, because of the logistics – the need for the custody officer to make an initial decision informed by the Gateway and refer the offender to the OMT at a different time and place – it made operational sense to separate consent into two parts: an initial agreement to return to consider a Turning Point contract; the formal meeting and agreement with the OMT. In contrast with Bethlehem, this approach to consent post-randomisation did not lead to high levels of attrition. The difference would appear to be that the incentive to participate – the opportunity to avoid formal prosecution and criminal conviction – was sufficiently substantial and immediate to encourage potential participants to attend the initial contract interview. There was also, in contrast to the delays that are inevitable in arranging a restorative justice conference, a very short, no more than 48-hour gap between the "pre-consent" in custody and the confirmation in the meeting with the OMT. Furthermore, once that had happened, it appeared, from exploratory interviews that the field researcher

conducted with the Stage 2 offenders, that offenders were generally engaged in the process. The subsequent analysis showed that breach rates were similar to those for non-attendance at court (DOC/TP/21).

4.11.1 (d) Covert Reassignments

Whilst they were important factors, it was, therefore, neither the ethical and legal issues nor victim and offender consent that that appear to have directly contributed to high levels of attrition in Stage 3 of Turning Point. The most significant contributor appears to have been what Dennis (1988) labelled "covert" reassignments. As we have seen above in the analysis of the case studies and prior pre-court diversion RCTs, such covert manipulations of the assignment are a predictable and substantial risk of bias and were a consistent feature in most of the earlier trials (for example, Lincoln et al., 1977, Klein, 1986, Byles and Maurice, 1979 and Little et al., 2004).

As Table 4.3 suggests, drawing on Braucht and Reichardt (1993), Sherman (2010) and Ariel et al. (2012), it is possible to identify a working hypothesis of the key steps that might be capable of mitigating these risks. The operational constraints of austerity and the police custody environment meant that one of these steps – entrusting the process of random assignment wholly to a research office – was not feasible in Turning Point. There was neither the research funding available for a 24-hour team nor was the project board persuaded to adopt such a model, which was inconsistent with the Chief Constable's strategy in favour of enhancing frontline discretion. Moreover, Braucht and Reichardt (1993) and Ariel et al. (2012) had suggested that this might not be necessary if a suitable computer-based randomiser could be implemented.

The Turning Point case study suggests that Braucht and Reichardt's (1993) optimistic reliance on technology is not borne out by experience. Ariel et al.'s (2012) more cautious acceptance of the technology allied to the personal support of the research team appears to come much closer to an effective model. However, it appears that even this model may be insufficient without

attention to the rigour and completion of the cycle of the tracking and feedback, for which Sherman (2013), Slothower et al. (2014) and Sherman et al. (2015) argue. Indeed, Turning Point suggests that the tracking and feedback may be more significant than the particular method of random assignment in a police RCT.

Stages 1-3 – Treatment Integrity 59%	**Stage 4 – Treatment Integrity 92%**
• Gateway randomising cases • Email alert to members of the project and research team • Monthly project board meeting to discuss progress • Internal (police) resource compiling a master spreadsheet on treatments • Custody Inspectors following up override decisions • Field researcher oversight (for part of Stage 2 and early part of Stage 3)	• Gateway randomising cases and refusing repeat custody numbers • Email alert to the project and research team • Monthly project board meetings to discuss progress • External (researcher) resource compiling master spreadsheet on cases from point of Gateway randomisation through to treatment completion/court (control) resolution • Custody Inspectors following up override decisions • Field researcher oversight, feedback and briefing

Table 4.4: Comparison of the Random assignment and tracking processes in Stages 1-3 and Stage 4:

4.11.1(e) Tracking

As Table 4.4 sets out, the differences of approach between Stages 1-3 and Stage 4 were relatively small details. They can be divided into two distinct types of activity: technical improvements to the randomiser to close identified loopholes and reduce the workload on the custody staff; tracking, training, briefing and feedback. Beyond the difference in the activity, there was also a difference in the proactivity of the project and research team to problem identification and resolution. In Stage 3, which saw considerable effort expended through the first six months on managing the expansion of the experiment into the two additional LPA's, the tracking process was largely passive, relying on the combination of the email notifications, spreadsheet and review meetings to alert the team to emerging issues. However, as the field researcher's review in February 2013 demonstrated (DOC/TP/17), this not only failed to reveal underlying problems, it gave a comforting impression of progress as the eligible cases mounted up. As the conclusions of the first interim report show, the research team, relying largely on the passive monitoring, thought that the experiment was on track to deliver the required sample in 2013 (Neyroud and Slothower, 2012).

In Stage 4, with the field researcher back in place full time and the addition of a research team data analyst, a much more active tracking process was developed. As soon as the email alerts arrived, cases were being checked, mistakes addressed and issues fed back in person and followed up with emails (DOC/TP/17). As such, the Stage 4 tracking process mirrored the treatment intervention it was tracking, which relied on celerity, particularly the speed of the process from custody to contract, as a key part of the change mechanism. Two subsequent Cambridge-led police field experiments have deployed a similar approach. Ariel used a team of student 'pracademics' to support treatment integrity in a multi-RCT Body Worn video experiment (Ariel et al., 2016b). Drawing on lessons from the earlier Rialto experiment (Ariel et al., 2014) and a number of other experiments, including Turning Point, Ariel ensured that a set of connected processes were in place: the field managers were all police staff who had been trained in experimental design and implementation; they had all agreed a "rigid trial protocol" and signed up their agency; they had a structured fortnightly feedback process on treatment

integrity of cases; they undertook detailed dip-sampling of Body Worn video metadata and cross tabulated it with random assignments; they communicated protocol breaches “immediately” to supervisors. A similar tight feedback loop was also developed for the Trinidad and Tobago Hotspots experiment (Sherman et al., 2014), where detailed monitoring of compliance with hotspot patrol dosage was fed back to supervisors in a regular “copstat” meeting.

4.11.1(f) Tracking and Professional Discretion

Turning Point, the multi-site Body Worn Video trial and Trinidad and Tobago RCT raise important questions as what is sufficient and effective tracking and what is the most effective balance between control to achieve compliance with random assignments and professional discretion to encourage ownership in police RCTs? In all three trials the key issue that presented itself was the need to balance compliance with an experimental protocol with professional discretion to determine eligibility, manage risks and carry out core preventive tasks. In both Operation Savvy (Ariel et al., 2015) and the Peterborough Hotspot RCT (Ariel et al., 2016a) it had become apparent that the staff, in this case Police Community Support Officers, were resistant to the tightly tasked patrol schedules required to deliver the preventive dosage. Groff et al. (2015) found similar problems with the more experienced patrol officers tasked in the second Philadelphia patrol experiment, in contrast to the relatively compliant “rookie” officers fresh from training who were used in the first Philadelphia experiment (Ratcliffe et al., 2011).

These experiments suggest that the process of tracking should be seen as a component of the much broader framework of leadership and management of the experiment, which we are going to explore in more detail below in Chapter 5. However, in respect of the specific issue of managing attrition, the Turning Point experiment suggests that tracking cases remotely through email alerts and monthly management meetings is insufficient, even where that process has been supported with training. Grol and Grimshaw (2003) highlighted the importance of adopting “multi-faceted interventions” in their systematic review

of the most effective ways of delivering implementation in interventions in frontline healthcare. On their own, they found that education and training, reminders, feedback on performance, computer decision support and small group meetings all produced small, but positive, effects, but that a combined approach was more effective than the individual strands. Grimshaw et al. (2012) suggested that the best results were to be achieved by an "evidence-based" tailoring of the multi-faceted approach to the particular challenges of the intervention. This suggests that, whilst it is possible to identify from the evidence the core components of an effective approach, a cycle of testing, tracking, revising and reforming the approach is a necessary path to more successful approaches.

In Turning Point, by Stage 4 and through the detailed review of lessons from Stages 1-3, the research team were in a position to draw on Grimshaw et al.'s (2012) advice and tailor both the Gateway and the training, briefing and feedback process for the field challenges of the experiment. Just as the "copstat" approach was developed in Trinidad and Tobago to respond to problems with dosage in the early stages of that experiment, so the revised Gateway Mark 3, personalised feedback and small group training and briefing in Stage 4 were developed in response to the lessons of Stages 1-3.

Three factors militated against the research team achieving that combination of approaches more quickly in Stage 3. Firstly, the decision to expand from 2 LPA's to 4 during the early part of Stage 4 meant that a small research team was severely stretched in delivering the additional training and support and building the relationship with and trust of the senior management teams in the new areas. Weisburd et al., 1993 have warned about the dangers of growing RCTs in order to achieve a case flow target, but they were focused narrowly on the risks of heterogeneity in the sample. Moreover, when Strang (2012) highlighted the challenges of sustaining the coalition to deliver an experiment, she was drawing attention to the risks to the existing coalition rather than the risks inherent in expanding the coalition to include partners and frontline staff who were not part of the original coalition. Birmingham West and Central and Birmingham North presented both different operational issues and two

contrasting management and Offender Management teams that took time and effort to absorb in to the experiment. This was doubly difficult as a result of the second issue. The temporary absence of the field researcher, as in Dunford et al.'s (1982) Memphis experiment, removed a key element of the active management of the day to day processes of the experiment.

The third and final issue was technology support. The author had substantial experience running three police forces' technology support and then the national body supporting the whole police service. That experience suggested that any project based technology support would, generally, take more time to implement given that small projects such as creating the Gateway or the decision support tools for treatments will tend to sit in a queue behind larger projects and only progress where they had a very senior sponsor with a budget. In this case, partly as a result of that concern about timescales, the initial version of the Gateway was designed in house by the research team. As such it benefited from neither the wider University technology support nor that of the force. Given, as we have set out above, the relative complexity of the Gateway that was developed for Stages 1-3, this was an obvious potential point of failure. There were relatively frequent complaints in the early part of the experiment about glitches with the Gateway. As a number of the Body Worn Video experiments have illustrated (Owens et al., 2014 and Grossmith et al., 2015), where the technology does not work smoothly there is a substantial risk that busy frontline staff like custody officers will either stop using it or look for work-arounds. The systematic review of Clinical Decision Support Systems (Kawamoto et al., 2005) reinforces this potential risk. In Turning Point, the lack of effective technology support was a persistent problem which was only eventually resolved by the out sourcing of the Gateway in Stage 4.

4.11.1 (g) Summary: Attrition (after random assignment):

As far as attrition after random assignment in Turning Point is concerned, the case study and analysis of lessons from other RCTs and the literature, would seem to suggest a set of key hypotheses for future police RCTs, particularly

where the research design is a trickle flow. The steps taken in Stages 1-3 (as set out in Appendix 5 and Table 4.3) were necessary but not sufficient to ensure a high level of treatment integrity. Resolving the ethical and legal issues was an essential part of the clearing the way to start the RCT. Piloting and pre-testing the eligibility screening and randomisation Gateway was absolutely necessary to build the tools and develop the research and project team and frontline staffs' understanding of the experiment. The training, in field support, tracking of eligibility overrides and system of email alerts and monthly meetings provided the basis for a model of training, tracking, feedback, support and monitoring but was not, of itself sufficient given the complexity of the experiment and the number of staff involved in its delivery. The Turning Point Gateway was a much more ambitious eligibility screening and randomisation tool than Braucht and Reichardt (1993) had envisaged in recommending a computer based solution to the problems of field randomisation. In fact, it was better described an algorithmic decision support tool than a simple random assignment tool. Sherman's (2010) very well evidenced concerns about the risks of attrition from entrusting frontline staff with the randomisation process proved to be, at least partly, realised as a result of a combination of problems with the technology support, over reliance on passive tracking and police custody officers continued exercise of discretion outside the bounds of the experimental protocol. As the comparison between the Consort for Stage 3 (Figure 4.2) and Stage 4 (Figure 4.3) demonstrate, the combination of addressing the technology support, revising the Gateway, providing active field support and active tracking of cases appears to have redressed those shortcomings sufficiently to achieve a low level of attrition.

4.11.2 Treatment Integrity issues: Treatment delivery:

The challenge of delivering the Turning Point treatment in the field had both common issues with attrition and some distinctive features. The progress towards higher levels of consistent treatment delivery followed a slightly different profile to the reductions in attrition. The discussion in this section is

organised as follows: the stepped approach; ensuring consistency of delivery; expansion and restarting; developing the treatments; the prescribing tool; the treatment delivery team; the trade-offs.

4.11.2 (a) The stepped approach:

As the details in Appendix 5 set out, there were six key steps towards ensuring treatments were delivered consisently: the pre-test of the initial treatment approach in Stage 2; the CPS assessment of the appropriateness of the treatments in Stage 2; the revised treatment model, training and guidance model rolled out in Stage 3; the creation of a coding model to assess the consistency of the Stage 3 treatments; the development of the "prescribing tool" during Stage 3 and its roll out in Stage 4; the addition of the LS-CMI needs and risk assessment model in Stage 4.

4.11.2 (b) Ensuring consistency of treatment

The central conundrum, to which these steps were directed, was the need to deliver both a consistent core treatment and a set of tailored conditions attached to it. As we have seen in reviewing the previous juvenile justice RCTs and the experience of implementing conditional cautions in England and Wales, delivering this combination of consistency and tailoring has proved problematic both in the experimental setting and in non-experimental operational implementation (Neyroud and Slothower, 2015).

The CRIMPORT (Appendix 4) was drafted (paragraph 8.1.1.) to provide clear criteria for assessing delivery based around the original theoretical conception for the experiment (Sherman and Neyroud, 2012). However, in a somewhat similar manner to that documented by Clarke and Cornish (1972), the treatment intervention evolved as the experiment developed. From the original, narrower focus on a "sword of Damocles" model of deferred prosecution with a limited set of conditions to encourage desistance, the focus gradually shifted to the quality and consistency of the conditions and their relationship to the best evidence. That shift required a more complex method

to assess treatment integrity. The original model only required evidence that the offender had agreed a contract which included the threat of breach. The "Turning Point" model meant that the research team had to develop a means of assessing the relevance, quality and dosage of every set of conditions (Slothower, 2014b). As such, the Turning Point experiment development presents a marked contrast to the experience of Feder et al. (2000), where the negotiation between the operational commanders and the researchers on the final treatment model produced such a low threshold for treatment delivery that the researchers concluded that the level of treatment was a probable reason for the lack of a significant effect.

4.11.2 (c) Expansion and restarting

One reason why Turning Point appears so different from the Portland experiment (Feder et al., 2000) may be the phased implementation and unplanned restart that extended the experiment well beyond its original timescales. Each of the phases provided an opportunity to revise the experiment, refresh the training and feedback the lessons of the last phase. It also provided an opportunity to feed progress upwards. Turning Point appears to be relatively unusual in producing two major interim reports (Neyroud and Slothower 2012 and 2013), which reported the hypothesis, progress and issues to the project board, senior management and senior researchers. Each Stage had specific lessons and developments. In Stage 2, the Gateway was revised and the preliminary exploration of the victim and offender experience, combined with the CPS review of the early cases provided learning that could be fed straight into the training for Stage 3. It also directly informed the emphasis on guidance to OMTs and YOS on the evidence on treatments. The issues with treatment compliance and consistency that arose in Stage 3 provided the stimulus to the OMT Sergeants, supported by the field researcher, to develop the "prescribing tool". The restarting of the experiment in Stage 4 allowed that innovation to have been pilot field tested and thus be available for full implementation for the final evaluation phase.

This model of multiple starting points was not wholly intentional, although a staged model had been envisaged from the beginning, when Stage 1-3 had been planned. However, the need to expand to four LPAs had not been fully factored into the conduct and management of Stage 3. It was disruptive and absorbed a considerable amount of the research and project team energy. Weisburd et al. (1993) have argued that expanding RCTs to secure a larger number of cases runs the risk of greater heterogeneity in the sample. In Turning Point, there does not appear to have been any such impact on the sample, but it did produce greater heterogeneity of the areas and the management teams involved and an increase in the numbers of staff directly involved in the experiment. This meant that the internal coalition of custody officers, OMTs, YOS, senior managers and project staff became more challenging to sustain through the experiment. Strang (2012) identified the importance of sustaining such a coalition, although the primary focus of her analysis was on the range of external partners that needed to kept on board. In Turning Point, strategic support, the peer group networks of OMT sergeants and the relationship between the project and research team all provided important mechanisms to hold that coalition together. We are going to explore these further below in Chapter 5.

4.11.2 (d) Developing the treatments

Although the experiment became more complex as the boundaries expanded and the treatment intervention became more sophisticated, the additional time which was required allowed the range of treatments to be developed and become well established. This was in marked contrast to the evaluation of the MoJ pilots (Ames et al., forthcoming) where many of the treatments were not ready when the pilot started and some were still not even in place as the pilot evaluation finished. Part of the problem was that there was a very truncated pre-test preparation phase. The universal feedback from the frontline staff surveyed was that insufficient time and effort had been allowed to implement the pilots properly. Groff et al (2015) encountered somewhat similar problems. They were testing the relative effectiveness of three hotspots treatments – targeted patrol, offender management and problem-oriented policing (POP).

Training the POP officers and putting the POP model in place took considerably longer than the other treatments. It is not completely clear from the evaluation how effectively POP was implemented and it is, therefore, difficult to draw any reliable conclusions about its relative effectiveness.

Groff et al. (2015) were, however, only testing three treatments, two of which, targeted patrol and offender management, were well understood and relatively established approaches. In pre-court experiments where the treatment intervention involves tailoring the intervention to the offenders needs and risk, whilst the core treatment may well be relatively specific, there are likely to be a number of potential variants. Experiments such as those managed by Lincoln et al. (1977) and Little et al. (2004) both involved a more complex array of a core treatment and tailored interventions. Neither was able to deliver a high level of treatment integrity, despite, in Little et al.'s case, a relatively small sample.

In conclusion, it is apparent that the more complex the treatments that are being tested, the longer and more thorough the dry run needs to be.

4.11.2 (e) The "Prescribing tool"

The challenge of securing consistent treatment proved equally significant in Turning Point. As we have documented above, the original model relying on professional discretion produced significant inconsistencies, which further training, guidance and tracking was able to ameliorate but not resolve. The levels of consistency achieved by the middle of Stage 3 were comparable to or even slightly better than those in Little et al.'s (2004) RCT. As Slothower's (2014b) analysis showed, it was the development of the "prescribing tool" that enabled a significant change to high levels of consistency and compliance. The tool was co-produced by the peer group of OMT sergeants with the support of the field researcher. As we have seen in the case of the Gateway, which could also be described as a frontline decision support tool, a computer based tool on its own will not deliver consistent decisions. It has to be embedded in and support a set of tracking and supervision routines that

reinforce the desired outcomes. The “prescribing tool” encouraged OMT staff setting the conditions to use drop down menus with pre-loaded interventions which had dosage and tracking requirements specified. The tool presented the data to the OMT supervisor who had previously relied on the submission of a paper record. The fact that the tool had been developed in house largely by the teams that were responsible for the work probably helped to overcome some of the early glitches. In contrast, the level of patience with problems experienced with early versions of the Gateway, which was seen very much as a requirement imposed by the research and project team, appears to have been far less.

The “prescribing tool” was a key development for which there is no other obvious comparable model amongst the police RCTs reviewed in Chapter 2. Indeed, in more than a decade’s experience in England and Wales with conditional cautioning, no such tool has been developed to support the process of setting conditions. The author was involved in the original design and development of the ACPO Matrix (ACPO, 2009), which was intended to support more consistent decision-making about out of court disposals in police custody centres. However, the Matrix has been inconsistently implemented (CJJI, 2009) and was designed to provide defensible decisions rather than decisions based on better evidence and consistent clinical practice. It was designed as paper based guidance and, whilst some forces did provide an electronic “look up” version to their custody officers, this was neither the general approach nor one that offered any support to the setting of conditions. It was quite apparent from debriefing the early conditions set in Stage 2 that even OMT staff, who were supposed to be managing offenders day to day as part of their core role, had little or no training or guidance available to support their setting of conditions. This appears to support the findings of a Criminal Justice Joint Inspectorate report on offender management which had identified that needs and risk assessment did not form part of the core competencies of the police officers working in OMTs (CJJI, 2014). The early conditions suggested that the OMTs in Turning Point relied quite heavily on the initial training that they had received as part of the project rather than any prior learning.

4.11.2 (f) The Treatment Delivery Team

One other mechanism for achieving consistency, which was strongly advanced by Sherman (2010) from his experience of the Restorative Justice RCTs, is the idea of a small specialist treatment team. We have seen how the Project Board decided not to adopt this model because of a combination of resource constraints and an operational preference in relying on the OMTs. On the face of it this would suggest that Turning Point provides evidence that the specialised treatment team is unnecessary. However, the reality of how the OMT's organised themselves to deliver the treatments suggests otherwise. All four OMTs tended to rely on a smaller group of officers within the team to do the Turning Point work. Moreover, the four sergeants formed themselves into a tight peer group who met regularly, discussed emerging issues and collectively problem-solved them. The Turning Point treatment team was, therefore, in effect, a small specialised team spread across the four LPAs with a collective management approach. In the replication of Turning Point, Operation Checkpoint in Durham, the specialised team has been adopted even more explicitly by entrusting the setting of conditions to a small, eight strong, team of "navigators" (Porter, 2015).

4.11.2 (g) The Trade-offs

Overall, the case study story of the Turning Point treatment delivery highlights the need to understand and test the trade-offs between the internal and external validity of the experiment. Shadish et al. (2002) discussed this in terms of the balance between efficacy and effectiveness studies. They suggested that "the internal validity of the inference that assignment to condition caused outcome does not require the treatment to be fully implemented." (p.319). Researchers conducting effectiveness trials, in this analysis, "recognise that treatments are often administered in the real world with less than full standardization and implementation" (p.319). As we have seen in the analysis of juvenile justice RCTs in this chapter, that dictum may have been stretched too far in many of those trials. The extent of this can

summarised by Dunford et al. (1982)'s conclusion that most of those studies were more of an illustration of the problems of pre-court diversion than a test of its effectiveness.

In Turning Point, the testing of the trade-offs and their impact was facilitated by the phases of the experiment and the opportunities that that provided to adjust and discuss progress. That process of adjustment presented another risk to the experiment: the extent to which the treatments delivered were sufficiently stable to be able to draw causal inferences with confidence across the experiment. Although it could be argued on a strict interpretation of the CRIMPORT that a high level of treatment delivery had already been delivered by Stage 2, the project board and research team decided that they were not happy with this benchmark and that decision provoked the subsequent development of the "prescribing tool". As we have seen, the combination of the tool and the process of training, tracking and feedback meant that the treatment being measured by Stage 4 had evolved significantly in its sophistication and consistency from the original Crimport specification.

As such, the way in which the experiment was run has many of the elements of Shepherd's "teaching police station", in which he had envisaged that clinical practice would be both tested and developed. Further support for this point is provided by fact that three of the members of the project board embarked on a Cambridge Master's degree during the experiment, with one completing a dissertation examining a key aspect of Turning Point (Hobday, 2015) and the other two undertaking experimental studies.

4.12 Conclusions: the lessons of delivering high levels of Treatment Integrity?

This chapter has explored three, linked research questions: the lessons for the conduct and management of police RCTs that can derived from published case studies and juvenile justice studies; the extent to which a case study of a high integrity experiment can add to those lessons; the extent to which the

lessons from both might add to or modify “novice theory”. The first question led to the development of a potential framework of operational factors that appeared to be important in effective conduct and management. The second, the case study, was used to develop and expand on those operational factors. Finally, taking the two together, the analysis concluded that, beyond the operational factors, there were some more strategic, “protective factors” that were also critical. These will be explored in the final chapter, Chapter 5.

Treatment integrity is central to the argument in favour of using experimental methods. Demonstrating a high level of treatment integrity is essential to the claim of any experimental design to be able to demonstrate a causal inference between the treatment intervention under study and the outcomes presented in the subsequent evaluation. Yet, the methods to achieve high levels of treatment integrity are generally accorded far less space in the key texts than the discussion of randomised design and the statistical analysis. As an illustration of this, the space devoted to treatment integrity in three seminal books on experimental designs can be summarised as follows: Shadish et al. (2002) provide six pages (pp. 314-320) on implementation in a textbook of more than 500 pages; Torgerson and Torgerson (2008) devote two paragraphs on pages 139-140 to the problems of attrition; Boruch (1997) has more generous treatment in his chapter on “operations” (pp. 164-195). Although Boruch comes closest, none of these texts provide anything like a comprehensive guide to achieving high levels of treatment integrity. Indeed, Shadish et al. (2002) devote more space to the relative merits of the statistical techniques to cope with the expected shortfall.

There are also very few texts that are uniquely devoted to the challenges of treatment integrity in experimental criminal justice studies, let alone in the specific field of policing. Indeed, until the more recent growth of case studies that we discussed above, only Dennis (1988 and 1990) and Sherman (1992 and 2010) had provided a detailed and empirically underpinned analysis of the issues. Yet, this case study of Turning Point set out in this chapter and other recent case studies (Sherman et al., 2014, Sherman et al., 2015, MacQueen and Bradford, 2016 and Famega et al., 2016) suggest that it is

very important for the rapidly growing field of police experimentation for us to understand the issues better and document them more transparently in published evaluations as the standards set out in Flay et al (2005) and Gottfredson et al. (2015) quite clearly require.

Significantly, a more detailed examination of the issues involved in treatment integrity begins to suggest that Eck may be wrong in his formulation of the "diabolical dilemma". Eck effectively argued that police researchers and practitioners could have either a high level of internal validity through a highly manipulated and controlled environment, or a high level of external validity by sacrificing control in favour of real world conditions. The analysis of Turning Point presented here suggests that the step by step process to achieve high levels of internal validity through high levels of treatment integrity can, instead, provide the evidence for a generalisable model of improved treatment in the real world. From algorithmic screening to the "prescribing tool", the methods developed in the field to underpin consistency and compliance in Turning Point have provided the basis for operational replication. As further evidence of this, Turning Point is already being replicated in Durham and Western Australia, where the lessons on treatment integrity have been carefully embedded.

All of which raises questions about the extent to which the findings of this case study on Turning Point can be generalised and how far they could add to or modify the "novice theory" discussed in Chapter 3. The approach set out in this Chapter has drawn on Yin's (2014) framework for case study research. The case study has drawn on multiple sources of evidence ranging from the Gateway records to the meeting reports and emails between the research and project board members and has included material drawn from the interviews with staff participating in the experiment. The analysis of juvenile justice RCTs and case studies and commentaries has been used to provide both a framework for analysing the case study material and supporting evidence for the emerging findings.

Yin's approach assumes that the research has been designed from the start as a case study and all the material has been gathered purposively against that research design. As has been set out above, that is not how this case study was developed. Instead, the material was gathered as part of the experimental design, management of the experiment and its evaluation. The original intention was to have used the material, where appropriate, to explain how the experiment was conducted. That means that this case study has not followed the systematic process from design and data collection through to analysis which is described by Yin. In practical terms data collection as part of the experiment preceded the design of a case study. Moreover, the case study ran in parallel with the search for and analysis of police RCTs and, as we will set out in Chapter 5, a wider exploration of the operational and strategic factors that may contribute to the effective conduct and management of police RCTs and, thereby, the treatment integrity of the experiment.

Both this evolving process of research and the distinctive features of Turning Point suggest that there are some limitations to the generalisability of the findings set out above. They apply most particularly to trickle flow experiments and, especially, to pre-court experiments – although Ariel et al.'s (2016a) study of multi-site Body Worn Video trials suggests that there is a strong read across into field trials of technology. Turning Point is very unusual in policing in that it was completely restarted as a result of serious problems with the intended evaluation phase. It was also unique up to that point in that the principal researcher, this author, was not just a serving police when the experiment was being developed but a serving Chief Officer. Yin's (2014) discussion of the skills required for case study work suggests that field knowledge is both a strength for providing an understanding the context and intervention under study and a source of potential bias in interpreting the findings. In particular, it would be surprising if the researcher's professional background, experience and curiosity has not led to different choices on the issues to be focused on.

As an example of this point, the author, who had been responsible for the strategy and management of police research focused on the challenges of

implementation. The latter became important because, as the experiment developed it became increasingly apparent to the research team that not only was the experiment yielding important insights on the conduct and management of experiments in policing, but also that there was a paucity of authoritative sources from which to draw for guidance and solutions. To some extent, this last fact may, at least partly, explain why novice RCT researchers, who are not, as Braga et al. (2014) found, already embedded in the global network of experienced experimentalists, appear to have problems with achieving high levels of integrity in their experiments.

This becomes even more significant when the rapid expansion of police-led RCTs becomes apparent. The list of in-flight police RCTs in Appendix 2 suggests that over 60% of the latest RCTs (in contrast to less than 5% before 2010) have a police officer or member of police staff as a principal investigator. Given the emergence of a new world of practitioner-led experimental research, it is important that the analysis in this Chapter has provided support for Sherman's (2010) grounded theory model for conducting and managing police RCTs. Sherman's model suggests a successful RCT, with a high level of treatment integrity needs attention to a wider set of factors than have been set out in this chapter, which has been very tightly focused on the mechanisms to manage attrition and ensure high levels of treatment delivery. Sherman's model explicitly assumed experienced researchers, not novices. This suggests that whilst novice theory provides an explanation for the pattern of treatment integrity that we described in Chapters 2 and 3, there are a set of "protective factors" and practices, which are also important and may be more critical when the three novice conditions apply. Feder et al. (2000) suggested as much in lamenting that they had not accessed the earlier literature. The analysis has started to draw out some of those wider "protective" factors – ownership, strategic support, education in the science of policing – which might be critical success factors. The final chapter of this research explores these protective factors and seeks to draw together the findings from this chapter and the opening chapters' analysis of the completed police RCTs.

5. “Learning to Experiment”: building a model for the conduct and management of high treatment integrity randomised controlled trials in policing

5.1. Introduction:

We have shown in the earlier chapters that the field of policing research is changing and there has been a significant expansion in the number of completed randomised controlled trials (RCT) in policing. Whilst the numbers are still very small compared to medicine, economics or education, 40 RCTs (Figures 2.5 and 2.6) were completed between 2010 and 2016 in contrast to 82 over the previous forty years. As a result, the average annual completion rate quadrupled from 2 to 8 per annum. The 63 “in-flight” RCTs (Appendix 2) suggest that this rate is continuing to accelerate. For comparison, Heneghan (2010) mapped the acceleration in medical research from less than 40 RCTs a year in the 1960s to more than 26,000 in 2010. He found that there were two key “landmark” moments – the passing of the 1000 per annum barrier in the 1970s and the 10,000 per annum barrier in the 1990s.

Whether the production of police RCTs has reached such a “landmark” may be too early to say. However, one key reason for the acceleration in RCT production in policing is a marked shift towards practitioner-led RCT research. In the first 40 years, there were only 4 RCTs published – less than 5% of the total - in which the listed authors or principal researchers were serving police staff (Earle, 1973; Sherman et al. 1992; Braga et al., 1999 and Shipley and Baranski, 2002). Between 2010 and 2016 that rate rose to more than 27% of the 40 studies. In the “in-flight” studies, that figure would appear to be over two-thirds (43 out of 63).

Whilst a significant proportion of that two thirds have undertaken their RCTs as part of a post-graduate qualification, there are also some signals of a wider dissemination of the approach. Most notably, Henderson and Magnusson, both Detective Inspectors, one in London and the other in Stockholm, completed RCTs outside the framework of formal academic supervision.

Magnusson embarked on a randomised field trial comparing two approaches to on-street warning of offenders for low level drugs possession after reading Sherman and Berk's (1984) Minneapolis experiment (personal communication, 2016). Both officers stated that they felt emboldened to experiment because they were aware of the larger conversation about "evidence-based policing" within their agencies (personal communications, 2016).

Whilst the number of RCTs in policing and the practitioner engagement in them has been growing, our analysis in Chapter 2 (Table 2.12) suggested that the level of treatment integrity may have declined in the most recent studies after a period of gradual improvement in the 1990's and early 2000's. Part of the reason for this may be, on the one hand, the exploration of new topics, such as Body Worn Video, which have, initially, raised new design and implementation challenges (Drover and Ariel, 2015 and Ariel et al., 2016a) and on the other, the involvement of new police forces in new countries, which have either had no prior experience of experimental research or have presented challenging contexts (Banerjee et al., 2012 and 2013).

However, given Braga et al.'s (2014) findings that most RCTs (up to 2012) had been conducted by a relatively small number of scholars within a tight, international experimental network, it seems reasonable to conclude that at least part of the issue with more recent RCTs may be involvement the "novice" experimentalists, more of whom are now police officers or police staff. This, in turn, suggests that the new discipline of evidence-based policing needs to be supported by better systematic knowledge of how to conduct successful RCTs with high levels of treatment integrity.

This chapter is concerned with exploring that question by drawing together the findings from Chapters 2, 3 and 4 and developing them further through an exploration of the experience of police officers in Operation Turning Point. The approach adopted seeks to develop a descriptive analysis of the protective factors at operational and strategic levels. The research questions that will be addressed are:

(c) To what extent is it possible to construct a model of the key protective factors that contribute to high levels of Treatment integrity in RCT field experiments in policing?

(d) **How might those factors add to or modify the "novice theory" proposed and analysed in Chapter 3?**

The Chapter is divided into four parts: an overview of the methodological approach; an initial overview of the existing literature on the conduct and management of police RCTs; the method, analysis and findings of interviews with a sample of police staff involved in Operation Turning Point; discussion of those findings and development of the model and its relationship with "novice theory".

5.2. A grounded theory approach to high treatment integrity in police RCTs:

A grounded theory study should be "designed to generate or discover a theory", which should consist of a "plausible relationship between concepts or sets of concepts" and can be reported "in a narrative framework or a set of propositions" (Urquhart, 2013:5). Glaser and Strauss (1967) described the approach as a practical rather than grand theoretical approach to developing theories from both qualitative and quantitative data. Grounded theory should not, they suggested, ignore the existing literature but should, as far as possible, start from a "non-committal" stance. For, Glaser and Strauss (1967) advise, the researcher cannot approach reality as a "tabula rasa" (p.3). However, they subsequently adopted divergent approaches. Glaser (1992) encouraged distance from the literature to allow findings to emerge from the research. Strauss (1987) accepted that this was often unachievable.

Glaser's distancing from the literature was not achievable in this research, where the overall approach has had three parallel and connected threads: an extended search for and review of all the published police RCTs; whilst at the same time the author has been managing a complex RCT in the field, which has formed the basis of the case study in Chapter 4; conducting interviews with the participants in the case study RCT.

Given the reality of field research, Urquhart (2013) suggests that it is often helpful to approach it in phases, which become progressively more engaged in the literature, moving from open minded exploration towards integration. It is Urquhart's phased approach that has been adopted in this chapter. Such an approach not only fitted better with the natural phasing in the research programme but also with the much longer research time frame dictated by a part time PhD. Instead of the more normal cycle of three years, usually divided into three sequential phases of literature review, field work and thesis writing, this research has evolved over nearly seven years from 2010-2017. This has both allowed Operation Turning Point to be completed (2011-2014) and for the initial focus of the research to broaden from a detailed focus on the evaluation of one RCT to the conduct and management of police RCTs more generally. As such, the 'research journey' itself has been a gradual, grounded appreciation of a significant and increasingly important gap in our knowledge about police RCTs.

In practical terms this has meant that there have been four connected and sometimes parallel phases: an initial engagement with the literature on experimentation and police RCTs, which was both necessary to frame the initial research proposal and to support the conduct and management of the Turning Point experiment; the search for and preliminary analysis of completed police RCTs; the development and deployment of an interview protocol in order to interview a sample of staff involved in Turning Point; detailed analysis of the treatment integrity in police RCTs and of the case study of Turning Point.

Mazerolle et al. (2014) suggested that there is an extensive literature on field implementation issues in RCTs, which we have drawn on in the earlier chapters above. There are important lessons to be drawn from this and the wider literature on implementation. However, there is significantly less literature and theory specifically on the conduct and management of high treatment integrity RCTs in policing, as oppose to the statistical approaches and methods of RCTs more generally (Sherman, 2010). Partly because of this, each of the phases of this research have taken a broadly grounded approach. In Chapter 3, the analysis of the data derived from the search for and analysis of 122 police RCTs was developed to suggest the importance of a 'novice theory' – that higher risk of bias from low treatment integrity is associated with novice experimenters, novice research stations and new topics. In Chapter 4, the analysis of the case study of Turning Point provided evidence for a model of the mechanisms required within a police experiment to ensure high integrity. This chapter will seek to use the data derived from interviews with Turning Point staff to explore the broader organizational and strategic eco-system and the protective factors which affect how those mechanisms are deployed to ensure an effective experiment is delivered.

5.3 Conducting and Managing police RCTs:

As Braga et al. (2014) showed, there have, until relatively recently, been relatively few experiments in policing from which to build a body of knowledge that we could use to develop a more generalized theory on the conduct and management of RCTs in policing. The search, which is set out in Chapter 2, has demonstrated that the body of RCTs is somewhat larger than Braga et al. found. Nevertheless, a heterogenous group of 122 Police RCTs still presents some significant challenges in developing such a generalised theory. Moreover, as Martinson (1974) and Dennis (1988) commented, some of the early experiments were both poorly documented and suffered from significant attrition and treatment delivery problems, meaning that the pool of high treatment integrity police RCTs has been even smaller until relatively recently.

This next section seeks to draw together the key elements of a model for conducting and managing police RCTs from the work of Boruch (1997), Sherman (2010), Strang (2012), Weisburd (Weisburd and Neyroud, 2011 and Famega et al., 2016) and Braga (2016). Boruch has been one of the key scholars promoting RCTs in social science. Sherman, Strang, Weisburd and Braga are four of the most important and prolific researchers who have completed police RCTs (Braga et al., 2014). After discussing Boruch and Sherman's models, the section will focus on police ownership of science, the role of pracademics and the importance of coalitions and partnerships.

5.3.1. Boruch's model

There is a substantial body of work on the conduct and management of experiments in fields such as medicine, education and social welfare (Boruch, 1997, Shadish et al., 2002, Torgerson and Torgerson, 2008). There are important lessons to be drawn from those fields for the implementation of police RCTs. Boruch (1997) drew those lessons together to provide an overview of the key "operations" tasks in an RCT. His sources included several police and criminal justice RCTs such as the Spouse Assault replication program, alongside education and social welfare examples.

Core Task	**Components**
Project management	Defining responsibilities and decision processes
Sponsorship	Identifying the problem, the treatments and the research objectives
Engaging Organisations	Engaging the target organization(s) and their leadership
Information	Developing a clear statement about the purpose and scope of the research

Research motivation (incentive)	Strategy for engaging the organization and its staff: interest in better answers; leadership support; stewardship (congruence of the research with the mission); precedent (prior RCTs); compensation (funding)
Credibility of Research partner	Experience, reputation and approach of researchers
Research site	Identifying a site with sufficient eligible cases, capacity and willingness to support data collection
Oversight and advisory groups	Oversight of the research and the external advisory input
Training	Training of researchers, data collectors and treatment providers
Termination	Process for determining whether the experiment has failed or has been completed and can be concluded

Table 5.1: Boruch's model of RCT Operations (derived from Boruch, 1997, Chapter 8)

Boruch's core operational tasks, which he described in Chapter 8, have been summarised in Table 5.1. This shows a set of strategic and operational activities that go well beyond the specific tactical mechanisms to achieve high treatment integrity that were discussed in Chapter 4. Boruch's framework makes some assumptions about the approach to RCT research. The most important is that the research will be directed and sponsored by a body outside the target organisation. The research itself would be carried out by a 'credible' academic research institution. In the way that he describes the process of securing a research site and support for the programme, there is little suggestion that the research is intended to be a shared enterprise.

Securing the engagement of the target organisation is driven by offering incentives, ranging from funding to answers to key questions.

Boruch's approach appears, therefore, to fit well with the 'outsider' model of police research (Brown, 1996). As an example of this, Dunford et al. (1982) and Dunford (1990)'s description of the process of initiation and engagement for the National Diversion programme appears to follow Boruch's steps. It was sponsored by the National Institute of Justice. The researchers were directed to several sites and sought to engage the agencies by persuading them that the NIJ funded programme would enable key questions to be answered about the effectiveness of diversion (Dunford, 1990). However, it may not be a coincidence that the research itself suffered from persistent problems with low treatment integrity because of problems with sample attrition from reassignments and failures in treatment delivery. Whilst the agency partners may have been engaged at senior level, courts and frontline staff appeared to feel little ownership and continued exercising their discretion outside the experimental protocol (Dunford, 1990).

There were some similar problems in the first domestic violence experiment in Minneapolis (Sherman and Berk, 1984), which had also largely followed the steps in Boruch's model. Sherman (1992) subsequently compared the experience between that first experiment in Minneapolis and the Milwaukee replication (Sherman et al., 1992). His analysis emphasised, above all, the importance of the frontline and middle management ownership of the experiment in the latter as a major contributory factor in securing high treatment integrity.

5.3.2. Sherman's model

Sherman (2010) incorporated that experience together with the lessons of a series of hotspot, offender management and restorative justice experiments to provide a framework for conducting and managing experiments that significantly moved beyond Boruch (1997). It was also much more tightly

focused on the challenges of implementing experiments in criminal justice generally and policing more specifically. A key difference from Boruch was that the approach centred on the research process as a 'contract' agreed between equals - the researchers and a participating agency - which started with the research question and the drafting of a research protocol. Sherman's key steps are summarised in Table 5.2:

Core Tasks	**Components**
Research question/motivation	Setting out the hypothesis and public benefit of the proposed research
The field station	Creating the initial partnership and agreement to research Developing the field station to support the "social foundations" of the research
The research protocol	A transparent (published) protocol for the research
Piloting and testing – a 'dry run'	Testing the design and, where necessary adjusting it
"Contracting" with key players	Recruiting professional participants Field Coordinator(s) Consulting key partners and groups Training key people
Well planned and responsive management of the experiment	Supplying cases Screening for eligibility Assigning treatments Delivering treatments consistently
Measurement and evaluation	Collecting data and analysing it
Publishing and communicating outcomes	Simple, core messages supported by the evidence

Table 5.2: Sherman's model of RCT Operations (derived from Sherman (2010)).

Sherman's model envisages a very different relationship between the researchers and the organisation studied and its staff. Rather than an externally driven process run by researchers on behalf of the sponsoring agency, the researcher is a 'partner' who needs to set out a sufficiently cogent case for the scientific and public benefit of the research to engage the agency providing the 'field station' in a jointly owned project from the start.

The concept of the "field station" is central to Sherman's argument. It is very different in nature and ambition from the "research site" of Boruch's model. Sherman (2010) derived the idea from the history and experience of experimentation and education in agriculture and medicine. He cited the precedent of agricultural research stations and teaching hospitals to support his advocacy of "field stations" where "experimenters" and "research partners" within the field agency could construct "firm social capital" through a "set of human relationships and social networks" (p.408). Experimenters and research partners needed to understand and sustain the network of "social elements" within the research station. Sherman identified that this necessarily encompassed five elements: the funders; the executive leadership of an operating agency; mid-level operating managers; those delivering the treatments; those providing the cases. For Sherman, each element needed careful and individual attention, tailored to their concerns and their power to enhance or obstruct the research, with the outcome of a high integrity RCT as the goal.

Despite the key differences from Boruch, Sherman also explicitly assumed an 'outsider-outsider' model of research. His 'contractual' model depended on active management by the principal investigators, who "will generally be PhD-level academics" (Sherman, 2010:415). He acknowledged Shepherd's (2003) advocacy of the practitioner-investigator model, but argued that those working in criminal justice had neither the time nor the investment in science to take on the key leadership role of principal investigator.

However, since 2010, through his involvement in the Cambridge Police Executive Programme (PEP), Sherman has helped to transform the landscape of police RCT research. From Police Chief Farrar's involvement in the Rialto RCT (Ariel et al., 2014) to the burgeoning numbers of in-flight RCTs, there are now eighteen RCTs in which Cambridge PEP students have been a principal or co-investigator. This list includes Operation Turning Point. Given this changing landscape of practitioner-led or involved experimental research, there is, therefore, a key question as to whether and in what ways Sherman's model might need to be adapted when the principal researcher is a practitioner or practitioners move from "research partners" to part of the research team.

5.3.3. Ownership of Science

A dimension of this question, which was not specifically explored in Sherman's model, but was central to Shepherd's practitioner-investigator model is the extent to which practitioner 'ownership' of the science and research in policing could be important in sustaining high integrity research. Shepherd (2002) had argued for practitioner academics in public services who would carry out research, act as informed advocates in the field and educate and influence their organisation. Shepherd (2003) lamented that there were at that time no police practitioner researchers. He recommended that a key step to remedy that deficit would be the creation of "police schools" modelled on the university based teaching hospitals, where education and research would come together with practice.

This would suggest a variation on Sherman's "field station" in which the "police school" rather than the specific research site would provide the base for experimentation. In a further exploration of the relationship between research and practice, Weisburd and Neyroud (2011) and Neyroud and Weisburd (2014a&b) argued that the police should take "ownership" of science. They deliberately used the term 'science' rather than research to

emphasise the broadest definition of science across policing. Taking ownership meant that the police would value science above untested experience. As a result, the police would take a central role in defining the priorities, supporting the production of science and actively engaging in the scientific process, including as researchers.

Weisburd (2005) had first drawn out the importance of 'ownership' from a case study analysis of two hotspot studies in Minneapolis (Sherman and Weisburd, 1995) and New Jersey (Weisburd and Green, 1995). He found that the "collaborative involvement of an individual able to utilize" hierarchical authority was critical to the implementation of a hotspots experiment (Weisburd, 2005: 241). In New Jersey, that individual, Captain Frank Gajewski, had not just been in authority, he had been "convinced of the failures of traditional approaches and the necessity of testing new ones" (p.237). For Weisburd, this emphasised the importance of Shepherd's model of integrated "clinical and research work" and the importance of practitioner belief in the science. However, what made that belief particularly impactive in New Jersey was that Gajewski could translate it into compliance with the trial requirements because of his "coercive power" within the hierarchy of the police. Weisburd went on to argue that the more complex the treatment was, the greater the need for such power to enforce compliance.

Weisburd's more recent analysis of treatment integrity in hotspots studies (Famega et al., 2016) has placed much less emphasis on hierarchy and coercive compliance and more on the "ownership" of the science. Famega et al. (2016) found that the difference between the treatment fidelity in three sites a multi-site study could, at least partly, be explained by the understanding of and valuing of the science. They stated that "when the police fail to take ownership, implementation tends to be shallow at best" (p.18). Ariel et al. (2016a) sought to leverage that ownership by using police officer "pracademics" as the field coordinators for a multi-site body worn video trial. They ascribed significant credit for the high levels of treatment integrity that were subsequently achieved to this innovation.

5.3.4. The role of "Pracademics"

Whilst Famega et al. (2016) and Ariel et al. (2016a) provide support for the importance of ownership, the debate between Huey and Mitchell (2016), Braga (2016) and Willis (2016) has illustrated that there are strengths and limitations to relying on the practitioner-academic or "pracademic" approach. Huey and Mitchell (2016) argued that the pracademic is uniquely positioned to be able to bridge the worlds of research and practice. Above all, in the context of a RCT, they could "employ their knowledge to convince, cajole and guide an organisation" (p.306). Braga (2016) emphasised the potential of pracademics to "develop the internal capacity" to meet "their own demands for scientific knowledge" and "strengthen connections to the academic community" (p.313).

In response, Willis (2016) suggested that the "pracademic" capacity to achieve this would depend on their position in the hierarchy and the extent to which the messages from the research were consistent with the received wisdom, "values and interests" of police leaders. For Willis, the chain of command, departmental culture and institutional context were all likely to provide boundaries on the type of questions and research that pracademics would be able to support. However, he conceded that RCT based research, focused on the effectiveness of treatments to reduce crime and improve public safety, might be more susceptible to pracademic engagement than other types of police research.

5.3.5. Coalitions and Partnerships

Braga (2016) touched on another strength of the pracademic model. He compared the analysis of police-academic partnerships carried out by Rojek et al. (2012) with the pracademic model. Braga argued that latter provided a more resilient bridge between research and practice than the variety of police-researcher partnerships described by Rojek et al. (2012). Braga drew on his own experience as an embedded criminologist in Lowell and Boston to

provide further evidence for his argument. Engel and Henderson (2015) endorsed Braga's position from their experience working with Cincinnati Police Department, where they found that the proximity of an embedded relationship between researcher and police was generally worth trading against independence.

Strang (2012) has focused more specifically on the importance of "coalitions", including police-researcher partnerships, as an essential component of sustaining a RCT. She described coalitions as time-limited alliances for a common purpose between parties – a research organisation and a police force - with differing goals and institutional frameworks. She found, from experience in conducting twelve restorative justice experiments, that, particularly with such trickle-flow designs which were likely to run for more than a year, the level of energy and attention required to sustain a research-police coalition was very significant. Strang identified three elements – the foundations (intellectual, social and legal), implementation mechanisms (which we have developed in Chapter 4) and leadership – as essential to maintaining the momentum and cooperation to complete a high integrity experiment.

The first two elements are common to Sherman (2010)'s model, but the additional emphasis on leadership is an important one. For Strang, the approach required needs to go beyond a single style. Transactional encouragement, tight management of tracking and, indeed, coercion to achieve compliance needed matching with transformational leadership, both to inspire participants and frame the research within the mission and values of the agency. Alongside leadership she also identified the importance of skills in negotiating difficult moments and rubbing points, which were inevitable over the course of the RCT.

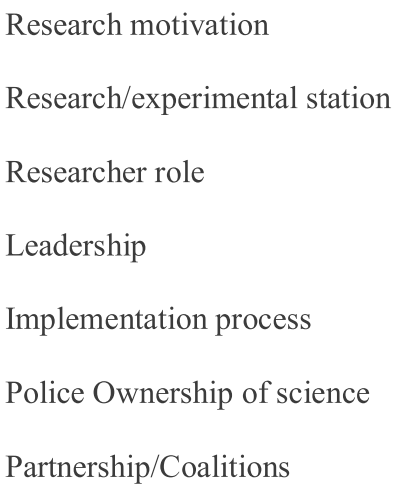


Figure 5.1: High level issues for the conduct and management of police RCTs

Drawing together the issues from the key studies in this field of conducting and managing field experiments in policing (notably Shepherd, 2003; Weisburd, 2005; Sherman, 2010; and Strang, 2010), Figure 5.1 seeks to set at a high level, those categories or constructs that appeared to provide an important starting point to explore further with the Turning Point participants.

5.4 The Turning Point interviews:

The Turning Point Project Board was keen to draw out lessons from Operation Turning Point: lessons for the force about initiating, supporting and conducting experiments; lessons from the process of implementation; lessons about the organisational learning from the experiment. This was felt to be particularly important given that, as we have set out in Chapter 4, the course of the experiment had not run as originally intended. The force was also committed to a strategy within which evidence-based approaches were a core strand. The research team, therefore, presented a proposal for a set of interviews with a sample of participants within the experiment. The project

board and, subsequently, the Chief Officer sponsor for Turning Point approved the approach.

The interviews were conducted in October 2013, in the middle of Stage 4 of the experiment. As we have demonstrated in the case study in Chapter 4, there was a significant change in treatment integrity between Stage 3 and Stage 4 of Operation Turning Point. The interviews were conducted six months into the final evaluation stage of the experiment at a point where many of the most problematic issues with treatment integrity had been overcome. Although it is now clear from analysing the treatment integrity that the early months of Stage 4 had provided a watershed between relatively poor treatment integrity and a stronger performance, the participants, including the researchers, would not have been in position to make that judgment at the time of interviews.

The high-level set of categories set out in Figure 5.1 provided the initial framework which was used to design the areas for exploration through theoretical sampling in the interviews (Urquhart, 2013). The original intention had been to code the RCTs in Braga et al. (2014)'s sample to build the initial framework, but, when this was attempted, the variability of reporting, particularly in the earlier published RCTs, meant that the level of detail provided on conduct and management of the studies was frequently insufficient to support reliable coding. Rather than using the RCTs at this stage of the approach, it became apparent that it was a more viable option to return to them and the other RCTs in the new sample of 122 in the discussion of the findings and development of the model.

5.4.1. The Sampling process

The sampling for the interviewing was determined by the need to focus on a purposeful sample of the most informed participants (Morse, 2007). The proposed list of interviewees was set around the key processes of the experiment – custody, offender management, project management and local

police area leadership. Except for the leadership and project board interviewees, where it was possible to make advanced appointments, the final list of custody and offender management staff was dependent on operational availability and business during the week agreed with the Project Board and local commanders for the research. The final list of interviewees was as follows:

Custody:	N = 5
Offender Management:	N = 7
Project Board/Strategic Leadership	N = 6
Total	N = 18

Miles and Huberman (1994:34) offer six general questions against which to judge a qualitative sampling plan: relevance to the conceptual framework and research questions; the likelihood that the sample will produce the phenomena the researcher is interested in; the likelihood that the sample will enhance the generalizability of the findings; the likelihood that believable descriptions will be produced; the feasibility of the sampling plan (cost, access and coding time); ethicality of the approach.

In this case, the operational considerations around abstracting custody and offender management staff for at least an hour for the interview needed to be balanced with the logistical implications of the transcription work and scale of the resultant data for the researchers. Guest et al. (2006) found that “data saturation” was achieved and issues were repeating themselves after twelve interviews. They also found that the main themes were established after six interviews. However, they cautioned that if the researchers needed to compare different groups or were dealing with very heterogeneous groups, more interviews might be necessary.

The project board agreed an interview schedule with a threshold of up to 20 interviews. The logistics in the week, particularly in custody, meant that the final count was 18. The research team concentrated, as far as possible, on securing six interviews with each of the three groups – custody, offender

management and management involved with the project board. The demands of custody meant that only 5 custody sergeants were able to be freed up. However, the project board members interviewed included two custody Inspectors. The author and the interviewer debriefed each day's sessions and made a high level note of issues emerging. By the final day of interviews, it was apparent that clear and consistent themes had emerged against the key themes designed in to the questions. Whilst comparison between the groups should be treated with some caution, the sample as a whole provided substantial data to support the research question and categories intended to be focused on.

5.4.2. Interview strategy and coding

The strategy for the interviews was designed to test and expand the data by exploration of the case study of Operation Turning Point. Morse (2007) encouraged the interview approach to be "targeted and efficient". The questioning approach, as set out in the interview protocol (Appendix 3), was designed to be both open ended at the start to ensure that the interviewee had an opportunity to put their own account, and then progressively more focused to ensure that the categories were covered. Inevitably, given that the interviews had to be accomplished within a live operational context, the sample, interview scope and timing were a compromise between the research needs and the operational constraints.

The interviews were conducted by Laura Bedford, a PhD Student from the University of Queensland, working to the interview protocol prepared by the author (Appendix 3). Her involvement was part of a research collaboration related to the process of learning from experiments. Furthermore, an independent interviewer was felt to be important for several reasons: the background of the principal researcher, as a recent former Chief Constable, provided a significant risk of bias in the interview process; the research team had also adopted an active role in the management of the experiment, the effect of which was an issue to be explored in interview; the importance to the

force of an external view. The Turning Point Project Board and the sponsoring Assistant Chief Constable approved the process and the independence of the interviewer.

The data derived from the interviews was transcribed and then coded using HyperResearch for Mac, which was developed as a computer aided coding programme (Hesse-Biber et al., 1991 and Lewins and Silver, 2007). The coding focused on "saturating" the categories in the initial coding and exploration as to whether the initial coding needed to be added to. In the next section, the findings from the interviews, the original high level categories have been used to organize the presentation of the material. The main findings of the section will then be drawn together with the findings from Chapters 2, 3 and 4 in the concluding discussion on the conduct and management of high treatment integrity experiments in policing.

5.5 The findings: conducting and managing the Turning Point experiment

5.5.1. Research motivation

The interviewees were asked three connected questions focused on the research motivation: why they thought that West Midlands Police had embarked on Turning Point; why the force had decided to undertake the trial as a RCT; how the experiment fitted with the West Midlands strategy. Across all three groups interviewed there were some consistent themes in the responses:

> "Well the strategy is always –the overarching strategy – to reduce crime. That is a very easy to understand strategy. How we reduce crime, I think… one of the strands is taking an EB approach to what works." (IV/TP/LPA1)

> "Because I think it is potentially there is a smarter use of resources, a better service to the public and to victims of crime." (IV/TP/CUS4)

> "to prevent offenders reoffending, save money, and keep them out of the CJS, and at the same time provide some satisfaction to victims." (IV/TP/OM7)

The respondents were generally clear that the experiment fitted both the force's strategy to reduce crime and improve services to victim and the financial context, in which the force was being required to make major savings. One offender manager was perhaps more cynical about this aspect of the trial, seeing it as "It is a very good way of saving money as well and that is what the government wants" (IV/TP/OM5).

Several respondents suggested that the force had had a past record in innovation with out of court disposals. They saw Turning Point as a natural extension of this and, particularly for the offender managers, consistent with the force's continuing investment, despite austerity, in the rehabilitation and management of offenders.

There was, generally, a positive recognition that, alongside the crime reduction goals within the strategy, the force's motivation was linked to a wider desire to innovate and adopt a more evidence-based approach:

> "We need to do more of what does work and we need to do less of what doesn't." (IV/TP/PB1)

> "I think we have both in the last 12-18 months, embraced a kind of EBP approach in a number of ways." (IV/TP/LPA2)

> "it is explicitly written down in actual fact, that we should take an EB approach" (IV/TP/LPA1)

In this context, the decision to use a RCT design was seen as part of the commitment to evidence-based policing. The "idea" for Turning Point was

seen as coming with the RCT design attached and the force was “not new to it, they had already spoken of an RCT happening before in WM” (IV/TP/LPA1). Indeed, West Midlands Police had already undertaken an RCT into targeting repeat anti-social behaviour, which had proved difficult to implement. The same respondent confirmed this: “the process of implementing a RCT is so difficult that it doesn’t sell itself.” That meant, in his view, that, for a RCT to be justified, the question had to be a very important one and the need for accurate results paramount.

In general, however, the respondents did seem both to understand the research question and were able to relate it to the force’s strategy and recent commitment to an evidence based approach. By October 2013, there had been at least four phases of training and briefing about the experiment, the research hypothesis and the research design. At the very least, the responses to questions about the research question would seem to suggest that the training and briefing had had an impact. However, the interviews suggested that the collective motivation for the research went beyond the research question about offender desistance policing. Several the respondents saw the force as an innovator and the experiment as an important contribution to a wider knowledge community, which by 2012 included the newly set up police professional body, the College of Policing (Neyroud, 2011). During the experiment one of the members of the project board set up a local branch of the Society for Evidence Based Policing (SEBP), which had, by the time of the interviews, already gathered more than 200 members.

5.5.2. The Research station

Such developments in West Midlands Police might suggest that the force was a fertile “research station”, on the way to developing into a “field station” and subsequent events have borne that out. Not only had the force already done one RCT with Cambridge, it embarked on a third – Operation Savvy (Ariel et al., 2015) – whilst Turning Point was running and a fourth, a Body Worn Video experiment (Ariel et al., 2016a), shortly after the completion of Turning Point

Stage 4. Since then, in collaboration with Cambridge and the Behavioural Insights Team, there have been a further six RCTs completed or initiated (Appendix 2) and a further six are in preparation (Murray, personal communication, 2016), establishing West Midlands internationally as one of the most prolific police RCT field stations (Appendices 1&2).

The interviewees themselves recognized that there was a qualitatively different relationship developing between researchers and practitioners:

> "we have got Cambridge here doing lots of different projects, it is starting to feel a bit more like a science lab". (IV/TP/LPA2)

> "TP is probably the first thing I have seen where we have taken an academic and evidence based approach to solving a problem". (IV/TP/OM2)

Sherman defined the distinctive aspects of a "field station" by the quality of the human relationships and social networks linking the researchers and practitioners. In Turning Point, a key aspect of this from the outset was a small group of Cambridge M.St. (PEP) graduates in key Assistant Chief Constable and Superintendent roles. During the experiment that group expanded to include two members of the project team and one of the Offender Management supervisors. Two subsequently conducted RCTs and one researched key aspects of Turning Point.

The relationship was such that the initial engagement for the experiment came not from a formal approach by either the force or the researchers, but as follows:

> "So good relationship with Cambridge University, understood the theory, spoke with [one of the principal investigators], [who] said well, why don't we try this [the Sword of Damocles] in WM? And yep if you can make it happen in WM that is great" (IV/TP/LPA1)

The same respondent emphasised the importance of the personal relationships between the lead researchers and key senior officers in the force. However, as Strang (2012) observed, the 'coalition' for such a research programme creates obligations and expectations:

> "If they leave chaos in their wake they have just burnt the laboratory down and so the legacy of an experiment must be beneficial to the host organization otherwise they will never go back to them" (IV/TP/LPA1)

The obligation to give benefits to the force as the experiment progressed was seen as critical by the force members. One of the offender management supervisors also emphasised the importance of the "bigger picture" in which the force was able, through Turning Point, to contribute to the development of national policy and support the College of Policing (IV/TP/OM1).

Linked with that was the perceived need from the force for lessons to fed back as the project developed. The force "can't wait three years to find out what the paper say that is published in the journal of criminology or whatever" (IV/TP/PB3). The researchers, unusually for such a project, produced, in addition to the Crimport, two interim reports, which documented the research hypothesis and supporting literature, design, early findings and emerging issues. The reports were presented to the project board, to the Chief Officer team and to the Police and Crime Commissioner.

The interim reports were also important for the funder of the research. Sherman (2010) identified that there could be a significant risk of tension in a RCT from the triangular relationship between the funder, evaluator and local force. This could be particularly problematic where the funder was a national government or government agency which was driving a policy agenda. Such tensions were very apparent in Ames et al.'s (forthcoming) evaluation of the Ministry of Justice pilots.

In Turning Point, the funds for the research came from Monument Trust, an independent trust, and there were few strings attached. Monument was

committed to finding better ways to reduce the use of prison, but was happy to leave the operational implementation to the force and the research team, with the one proviso that the research had to be made available for the greater good beyond West Midlands. This allowed the researchers and the project board to be able to reassure even the potentially cynical amongst the participants that the research motivation was being driven locally rather than being directed from the national centre.

There was a downside to the independent funding in the view of one member of the project board:

> "TP is an experiment which Cambridge can fund. We are not funding this. We are happy for you to use us as a petri dish but no more. But at the same time they have said that they would like some results out of this. So we would like for example; to get an interim report that tells us whether there is any mileage in this, whether the early results are good." (IV/TP/PB4)

For the frontline officers in custody and offender management, the most obvious manifestation of the "research station" was the series of training and briefing events. The first of these, the two-day session at Cambridge was remembered with affection: "they could have more dinners down in Cambridge" was a quip from one offender manager who was asked what more the research team could have done (IV/TP/OM1). We have already identified in Chapter 4 that the symbolic effort of taking so many operational officers to Cambridge for the training had convinced some that the programme was serious.

However, the research and project team maintained a series of training sessions, seminars and Turning Point events throughout the programme:

> "They have provided people with training days, you know, they have regularly come round to many of the LPUs." (IV/TP/OM1)

The custody interviewees referred to the briefings and hands on support in the custody suites to sort out problems with the Gateway, the offender managers to the work on the ‘prescribing tool’ to support treatment. Their experience highlighted a very active, hands on role adopted by the research team.

5.5.3. Researcher role

For a project of this scale, the research team was small. The author was a principal investigator and research manager and “novice” researcher. The senior researcher, Professor Lawrence Sherman, was both a principal investigator and supervisor as well as the senior Cambridge presence. As we have seen, Sherman was one of the most experienced police RCT researchers. Dr Barak Ariel, from Cambridge, helped with the design and the Gateway. He had been involved, with Sherman, on the earlier West Midlands experiment on Anti-social behaviour. The field researcher, Molly Slothower, came onto the team from February to August 2012 and then back again from March 2013 onwards. She was also a “novice” researcher at this stage. Finally, there was the data analyst, Eleanor Neyroud, who started in Stage 4.

We have already demonstrated the importance of the field researcher in the case study analysis in Chapter 4. In this section, we want to focus more on the ways in which the researcher role was perceived by the interviewees and how that might have impacted on the integrity of the experiment. It was particularly important for this section that the interviews were carried out independently. The interviewer only knew the author from the research team and was, deliberately, in order to reduce the potential for bias, only provided with a high-level briefing about the project, including a copy of the Crimport and slides from the training sessions.

The questions designed to explore the research team role and contribution were headed by a very open invitation to discuss “the most important things that the research team had done to make the experiment work?”, with a follow up seeking things that could have been done better.

The first factor that was raised by several respondents was the importance of the author's "pracademic" status:

> His "influence on it is easy to underestimate, because he has, better than anyone I have seen, he bridges that gap between the police and academia. He has a foot hold in both camps". (IV/TP/PB2)

On other hand, the police officers also valued the field researcher's approach:

> who got down on ground level, was obsessed with getting it right, when there was a problem taking other people with her, changing their attitudes, investing in time and being available at the end of a telephone all the time – so everybody knew her and they knew what it was about and she had the ability to convert a cynic to convert. (IV/TP/LPA1)

> There was no, I hadn't sort of had any explanations as to why I was doing it. Obviously that they have explained that and sort of any sort of questions that I have got [the field researcher] is the person that you phone. Even if it 3 o clock in the morning – I wouldn't do that- but she has said" (IV/TP/CUS1)

They also recognized that they were getting much more than support for the Turning Point project:

> "she has managed to give a lot more to the force not just TP as well. So having a Cambridge based academic within the force has had a lot of spin off advantages to the force, that it probably doesn't even realize" (IV/TP/PB2)

This seems to support Sherman (2010) and Strang (2012)'s observation that the research team has a key role in building not just social but intellectual capital. Building and maintaining the social capital was, however, the main

task and one that absorbed most of the research team's time and energy. Sherman identified four key groups of people: agency executive; operating liaison; treatment deliverers; case providers.

In Turning Point, these groups were: the Chief Officers, who had also designated one Chief Officer as the corporate sponsor; the Local Area Commanders and Project board members; the offender managers; the custody staff. The research team were stretched to provide sufficient cover to all these groups, particularly in the hiatus during which the field researcher was away. Several respondents commented that too much had had to be done by email as a result.

There were a number of specific research team interventions that were mentioned when the interviewees as the most important things to make the experiment work. First and foremost, a number of the respondents, in all three groups, emphasised the importance of

> "the humility and the listening ear to spend time with people involved in the experiment at ground level." (IV/TP/LPA1)

The project board respondents valued the commitment to meetings, both formal project board and:

> "coming here very regularly so lots of those types of meetings, and then after a while the governance board kicked in and ok lets manage this using a project structure" (IV/TP/LPA1)

For the offender managers and custody officers, it was the time taken to explain the project, often in the workplace rather than the formal training sessions.

> "[the research team] has come and explained to me why we are doing it, so part of my feeling obviously comes from that input." (IV/TP/OM2)

They were also energized by the enthusiasm of the senior researcher, both in the initial training and in the February 2012 review day:

> the way he puts the whole project across really – persuasive is not really the right word (IV/TP/CUS2)

Above all, the frontline staff appeared to value an approach which was engaged, active and in their own workplace:

> "[He was very] much hands on in delivering some of the training to the staff. He has been very visible". (IV/TP/CUS2)

One of the reasons why this was possible was that West Midlands Police accorded both the author and the field researcher complete, unfettered access to police stations, custody suites and Headquarters buildings. They also provided the field researcher and the research team data analyst with full access to the force data systems for custody, crime and court results. This meant that the research team, with access guaranteed by West Midlands badges, could be present at any time and their contribution was not confined to pre-booked meetings and formal set piece events. It undoubtedly helped that the author was a recently retired Chief police officer, but even so, it is unusual for any researchers, whatever their prior status to be granted such open access. This was a concrete manifestation of the very strong, senior leadership support that the programme enjoyed.

5.5.4. Leadership

The custody and offender management interviewees were in no doubt about the very senior support for Turning Point. One of the offender management Sergeants put it bluntly:

> "I would say now fully committed. It is pretty easy to answer that one. The Chief supports it, therefore everybody else supports it. Simple as that." (IV/TP/OM7).

He went on to suggest that the ability to 'wave' the Chief's support at middle managers had been helpful:

> "it has been nice to say ... to line managers that this has the support of Chief Constable downwards were supporting this scheme" (IV/TP/OM7)

One of the project board members had a similar view, that in trying to get things done to support the experiment:

> "The high level command was inevitably very important. So have support from your senior ranks – CC, ACC and so on. Having support at Chief Superintendent level was very important and they have effectively told people involved that "you will do this". That is quite critical and that makes a difference." (IV/TP/PB4)

This appears to correlate with Weisburd's (2005) observations on the importance of key individuals with coercive power in a police hierarchy. However, a much subtler 'permission' to innovate was also seen as very important in the day to day management and problem-solving. Another of the offender management supervisors commented that:

> "We have been allowed to develop it and we have developed it to suit us" (IV/TP/OM6)

All the OMT supervisors recognized that they had been given wide discretion, time to meet and allowed to solve the problems, such as the variation in treatment integrity, for themselves, supported by the field researcher. One of the project board members commented that when the initiation of a significant

change to the programme had been proposed, he realized that he had permission to get on with it without referring it upwards.

In custody, the role of the Custody Inspectors, who had responsibility for the five custody centres was seen as central and, despite their focus on tracking custody officer decision-making, largely supportive:

> "I am really fortunate in that my direct manager…. was obviously very committed to the project and believed in it." (IV/TP/CUS4).

Indeed, matching Sherman's findings in Milwaukee, the interviewees emphasised the importance of a group of operational managers and supervisors in making the experiment work:

> "There are key managers – mainly supervisors who are keen to make it work. At quite an operational level there are some key players. Who are very keen to make it work". (IV/TP/PB2).

There were, however, frustrations and these were particularly linked to commissioning actions from other parts of the organisation, such as the ICT department, or securing funding for improvements to the Gateway or the setting up of the Victim's team. The ICT response was a constant issue. Turning Point, whilst a large experiment, was a small project within the larger change strategy that the force was pursuing as a result of funding cuts. The inability to secure timely support from the ICT department meant that Turning Point had to rely on bespoke, externally sourced support, which was expensive and did not always provide scalable and stable solutions. Early troubles with the stability of the Gateway were a case in point.

Overall, the author, whilst explaining the experiment and seeking to persuade middle and senior managers to support it, observed three broad responses: an enthusiasm for evidence-based policing; a career-oriented recognition that EBP and the experiment enjoyed the Chief's support; an acceptance, without enthusiasm or commitment, that Turning Point was a project endorsed by the

force. Overt opposition to the experiment was rare – the author only encountered one senior manager who expressed strong opposition to a RCT based on an ethical and legal objection to what he perceived was differential treatment.

Over the course of the experiment, it appeared that there were more enthusiasts and fewer sceptics. One of the OMT supervisors suggested that "some of the line managers who were a little sceptical at the beginning, can see the benefits of it" (IV/TP/OM7). One of the commanders described their own personal change of view as follows:

> I was quite anti-it to begin with… as in 'Oh it is an experiment' I don't really like the sound of the kind of ethical grounding to it, it sounds as if we are just furthering the studies of Cambridge or what else… I think I have a far more open minded approach to EBP, and I think that has come from TP really…. So I think on a personal basis it has opened my eyes and made me far more receptive to doing experiments, and so on and so forth…So I am perhaps a little bit closer to that and seeing the benefits that it can bring. (IV/TP/LPA2)

5.5.5. Implementation

Given that the experiment had failed to reach an acceptable level of treatment integrity after more than two years' work, it would have been understandable if the level of scepticism had grown rather than diminished. We have described the detailed steps towards delivering a high integrity study in Chapter 4, but there were some broader lessons on implementation that emerge from the interviewees' responses. The most important of these was the emphasis that all three groups put on the importance of taking time to get things right, and in doing so, listening, adjusting and revising the experiment step by step:

> "the test within a test to make sure we were getting it right, and I think that is bang on right. That is right. Because they would have flawed the results.

> Our plans weren't good, we were getting things wrong, we as OM were getting things wrong, the custody block were getting things wrong" (IV/TP/OM6)

> "This implementation is always a challenge that manager's face, and that therefore staged implementation makes perfect sense. I just think even before the stages that they did do, there would be benefits of a pre phase of greater planning or testing the ground." (IV/TP/PB4)

These comments parallel the feedback to the evaluators in the Ministry of Justice pilots (Ames et al, forthcoming), where the major issue raised was the wholly insufficient time to prepare for the trial and lack of adjustment as problems emerged. In Turning Point, it was not just that a step by step process was important for the process of implementation, a number of interviewees identified that this was an important part of the learning process:

> "It has to be in stages because it is a learning process all the way through". (IV/TP/OM3)

The interviewees also recognized the constant tension between a tightly imposed programme managed from the centre by a dedicated team and allowing staff discretion and space to solve problems. Overall, whilst raising concerns that the programme team had been too light on resources – a point that the project manager stressed particularly – there was considerable enthusiasm for the leadership approach that explained, encouraged and enabled rather that a more compliance approach:

> "there is no point in having pressed people, and so there was an element of persuasion –this is why you are doing it, and you are going to do it anyway so…" (IV/TP/LPA1)

> "People have been allowed to be creative. So that is probably the key, all that waffle I have just come out with, that we Police Officers have been

given the freedom to be creative with our idea and to try them out." (IV/TP/OM1).

There was, alongside this an acceptance of the importance of continued tracking of compliance with the experiment. Even by October 2013, there were:

"clear attempts to break that [the Gateway] and subvert it, as we found this morning with one particular case, and it gets picked up and it will get fed back to them and the useful bit there is being able to feed back to people whose Q14 kinda model…" (IV/TP/PB4)

and the custody staff accepted this:

"I have had one email off him [the custody inspector] saying "this custody record wasn't put through the randomizer", which could have been a positive, could have been a negative sort of thing. So it was good that he was checking and there was that control measure was there. To ensure it was being complied with" (IV/TP/CUS1)

The managers and members of the project board emphasised the importance of the formal foundations of the experiment – meetings, the discipline of programme management and the project team:

Plenty of meetings obviously, we used to have regular monthly meetings around different cases initially, how the project was working, the by project was working (IV/TP/CUS2)

One aspect of the way that these meetings were managed in Turning Point was seen as particularly helpful. The project board and the sub-meetings were deliberately rotated around the four Local Policing Areas, which allowed them to get "all the OM together and reps and Custody and youth services" (IV/TP/LPA1) and helped to create a sense of a bigger "Turning Point team"

across the whole of Birmingham. It was these type of meetings that could be used to feedback progress:

> "we have regular board meetings as part of TP where we feed back in things we have learnt, what has been happening that has been really good, success stories and what has not gone so well. So all the lessons we have learnt have been captured over a period of time." (IV/TP/OM1)

5.5.6. Police ownership of science

The continuing feedback as the experiment progressed seems to have been extremely important in engendering more than just compliance:

> "This isnt' just something we want you to do because someone upstairs has told us to do it, this is a project and these are the reasons behind why we want to do it. A rationale makes it so much easier to put your effort and your passion into, and you know why you are doing it." (IV/TP/OM2)

Part of that enthusiasm came from the personal experience of observing the individual effects of Turning Point's impact on offenders that the officers were dealing with:

> "It is brilliant. When you get that and they go off… they get jobs, they sort out their money issues, sort out their minor drug issues and get jobs… All things like that are brilliant and absolutely fantastic." (IV/TP/OM5).

> "It makes me think a lot more deeply about likely outcomes" (IV/TP/CUS4).

It was also very clear that the interviewees understood that they were undertaking a very different, scientific enterprise:

> “I am not an academic, I am just a thick policeman. But my understanding of randomized testing is that it is a good way of determining whether something actually works as opposed to and removing all of the extraneous things that might affect results and conclusions (IV/TP/CUS4)

> “We are interested in research for the sake of changing practice for the better.” (IV/TP/PB4)

One OMT supervisor, who had joined the experiment at the third stage, found that being part of the experiment had encouraged him to feel part of a wider professional community:

> “we are invited to the society for EBP. So what works been going on around the world really. And how much thought and consideration is put into these experiments and the reasons behind it.” (IV/TP/OM7).

For some of the interviewees, the experience was personally transforming:

> “I have learnt what an academic trial is… I have learnt some important management lessons on how to get this over to reluctant members of staff… how to present it. I have built up presentation skills. And I have got a real commitment now to EBP, to such an extent that I am hopeful of doing a masters at Cambridge now” (IV/TP/PB3).

As the experiment progressed a network of evidence based innovators developed amongst the leadership of the Local Police Areas. Initially, there had been one commander who had been the key advocate of experimental approaches and a national spokesperson for evidence-based policing. By October 2013, he was able to point not just to others who were enthusiastic, but others who were actively supporting other experiments (Operation Savvy):

> “the great thing here is you see ideas spread” (IV/TP/LPA1).

Importantly, it was not just amongst the managers that the ideas permeated:

> "Custody Officers particularly older Custody Officers such as myself are generally regarded as probably being the most cynical bunch of Police Officers that you will ever come across, and the most difficult to convince that radical change is every going to be a good thing. But very very few Custody Officers failed to see the potential benefits of TP." (IV/TP/CUS4)

Indeed, by October 2013, one commander, who had had doubts in the early stages, stated that:

> "I just think now particularly here it is accepted practice" (IV/TP/LPA2)

5.5.7. Partnerships and Coalitions

Getting to that point and sustaining the internal and external coalition necessary to sustain the experiment had entailed some important decisions from the outset. Reflecting on the problems encountered in experiments which had been reliant on multi-agency models (Dunford et al., 1982, Giblin, 2002 and Little et al., 2005 for example), the Turning Point model had been deliberately designed to ensure that a single agency – the police – made as many as possible of the major decisions about assignment and treatments.

The only exception to this was the involvement of the Youth Offending Service in delivering the treatments to young offenders. Unfortunately, it was not possible to secure permission to interview YOS staff within the timetable for the interviews. From observations and meetings with the YOS team and managers, it was apparent that, whilst they were initially unhappy to be left out of the original conception of the project, the Turning Point approach fitted well into their own view of their work. The YOS team were under severe financial and performance scrutiny because of cuts to the local government budget. Despite the additional work that Turning Point brought them, their manager consistently emphasised the opportunity that the experiment offered them to

demonstrate their effectiveness (personal communication). However, more could and should have been done, at an earlier stage, to engage the YOS who were “quite isolated from the police” and “there was a lack of interaction and that has proved difficult at times” (IV/TP/PB4).

However, the most important coalition for the experiment was the police-research partnership. The fact that the relationship was sustained through the crisis of discovering low treatment integrity, despite two years’ work, suggests that it had built very considerable social and human capital by 2013. As we have seen, part of that was due to the pracademics bridging research and practice (in both the research and, increasingly as more became Masters students, the project team). Partly, it was the hands on approach and level of commitment to the project – no one who had attended the initial Cambridge session had forgotten the experience. As one Custody Officer put it:

> “The passion for the project was clear from the outset from the people who were heading it up. And that passion was quite easily transferred to the people who were giving the input.” (IV/TP/CUS4).

A further important factor was the commitment of the research team to provide findings and lessons from the research even from the earliest stages:

> “I want some results. I don’t want the full results. I don’t care what the full results are, but I need to keep these guys interested” (IV/TP/PB3).

Some of these ‘results’ were provided at a very early stage through the combination of the CPS review of the Custody Officer screening and OMT treatment plans. Custody Officers were given feedback that the CPS view that their decisions were consistent. OMT’s were individually briefed about the issues emerging from the treatment conditions and responded by asking for more support about the evidence and further training.

That same process of initial results produced an important additional aspect of the research – the second and parallel RCT on victim satisfaction with Turning

Point compared to a court prosecution. This arose from the field researcher's exploratory survey of victims and offenders from Stage 2 which identified some significant issues with victim satisfaction. The results were used by the research team to propose a different method of engaging with victims, the effectiveness of which was tested in the second RCT (Slothower, 2014a). The results of that experiment – that Turning Point with the new victim's approach produced significantly better satisfaction than a court prosecution – provided the force with a major finding at a time when the Police and Crime Commissioner was searching for better support for victims.

With significant findings emerging and being shared in presentations and the two interim reports, the force and, in particular, the project team were able to hold a number of events to showcase the experiment: a SEBP event which attracted more than 200 attendees from surrounding forces; a national seminar with the national Police Chiefs lead and an Australian Police Commissioner in attendance; a West Midlands stakeholder event for agency partners and treatment delivery agencies. Key members of the project board secured support from the force and Cambridge to undertake a Master's degree and a number were promoted. Four of the Local Police Area Commanders involved have already gone on to achieve Chief Officer rank. In short, there was organizational, personal and professional gain from being a part of Turning Point.

5.6. Discussion of Findings: Towards a modified theory on conducting and managing field experiments in policing?

The discussion which follows starts by reviewing the strengths and limitations of the evidence presented in this Chapter. The second part draws together the factors that have emerged from the findings in this chapter, the mechanisms which we drew out from the case study in Chapter 4 and the "novice theory" that was examined in Chapter 3. The third part then focuses on the seven themes that we set out above linked to the three novice categories:

- research topic – research motivation and implementation.
- researcher – the researcher role and “pracademics”;
- research station – building a “field station”, leadership, partnerships and coalitions and ownership of science.

5.6.1. Strengths and Limitations

There are several limitations to the generalisability of these findings. Firstly, looking across the wider sample of 122 police involved RCTs, Turning Point appears to have been an unusual experiment not only because of its scale and duration, but also because of the multiple starting points and the complete restarting of the experiment. It was also a trickle flow experiment and the findings may well be less relevant to other designs.

Secondly, although the experiment covered a large part of the force area – the whole of Birmingham, the interviewees constituted only a small group (18) of the active participants (in total around 150 staff) in the experiment and do not include the much larger group of operational officers and managers who were not engaged in the experiment. We have relied on Guest et al. (2006)’s findings on the size of reliable purposive sample, but recognize we need to be conscious of the potential bias from such a sample. As we have set out, the final group was partly selected by the research team and partly by their operational availability during the week selected for the interviews. Around half the sample were supervisors or middle managers, some of whom had invested significantly in time and reputation in the experiment. However, it would also be fair to say that, at the time the interviews were conducted, participants did not know whether the experiment would, ultimately, be regarded as “successful” or not.

The interviews were only conducted at a single point in time – October 2013 – rather than in two or more phases, for example pre-test and post-test. A series of interviews at different stages of the test could have provided a stronger basis for drawing conclusions about the changes that were

happening during the experiment. However, the interviews were not part of the original design of the experiment and the need for them to explore the lessons of the experiment only became compelling as the challenges of implementation became clearer.

If only one set of interviews was to be conducted, focused on the questions of conduct and management of the experiment, October 2013 was the optimum moment to do so. Stage 4, where the treatment integrity had improved significantly was well established. The end of the experiment was not yet in sight. The research team were still projecting that the experiment needed to run for at least a further six months. As such, the experiment was both live and at its highest integrity after a period, in Stage 3, where there had been problems.

5.6.2. Building a model from the findings

In Table 5.4 we have drawn out the main themes that emerge from the interviews against each of the original, high level categories from Figure 5.1:

Category	**Narrative themes**
Research Motivation	• Research consistent with the police agency's mission and goals • A Police agency with an ethos of innovation • Briefing and training to explain the research and support the linkage with the agency mission and goals • Allowing police officers and staff to develop their own personal narrative of the research and the

	links to their own professional and organizational goals
Research Station	• Prior scientific research involvement by the agency • Importance of a small group of "connectors" able to link the research team and enablers within the agency • Establishing the social relationship between the researchers and operational staff through "ice-breaking" or "kick-off" events (the Cambridge sessions), training and meetings • Building the formal relationships of meetings, contracts and reviews

Researcher role	• Pracademic role as a bridge between research and practice
	• Field researcher focus on one to one and small group briefings to explain, problem solve and build confidence in the experiment
	• Researchers providing added value to the agency beyond the specific research programme
	• Researcher approach guided by:
	  ○ Listening and humility
	  ○ Explanation
	  ○ Enthusiasm
	  ○ Practical and active engagement
	• Open access for the researchers

Leadership	• Chief Officer commitment • Empowerment of the frontline to innovate within boundaries of the experiment • Key middle managers with commitment and capability to use positional power • Operational managers group responsible for project management and problem-solving, committed to the experiment • Allowing time for sceptics to 'cross the bridge' to support for the experiment
Implementation process	• Pre-testing, piloting and rolling out in steps • Learning 'as you go' • Persuasion rather than coercion within a framework of clear direction • Tracking key processes, fairly and with constructive feedback • Structured programme of meetings to track, review and problem solve • Using the meetings as an opportunity to engage officers participating
Police ownership of science	• Personal experience of the change process from the experimental treatment

	• 'Police science' education to build understanding of the experimental process and the hypothesis • Building a 'science community' within the agency and linking to wider national and international community • Providing opportunities for personal growth and learning • Expanding the small group of police science 'connectors'
Partnerships and coalitions	• Engaging multi-agency partners at the earliest opportunity • Single agency decision making and accountability wherever possible • Creating passion and excitement about the innovation being tested • "results" and feedback from the earliest stages of the experiment • Ensuring professional and personal benefits from the experiment

Table 5.4: Categories and narrative themes derived from the Turning Point interviews

5.6.3. Integrating and modifying "novice theory"

Table 5.4 has been presented as a set of narrative themes drawn from the analysis of the interview findings (Urquhart, 2013). As such, it summarises the findings from the interviews, which were focused on operational and strategic

factors. We have already established more detailed findings about the impact of novice researchers, new research stations and novel topics in Chapter 3 and key mechanisms or “implementation components” (Fixsen et al., 2005) related to maintaining high treatment integrity in Chapter 4.

In seeking a way to represent the linkages between all these elements, we have drawn on the model developed by Fixsen et al. (2005) from their systematic review of the implementation of social programmes. Whilst their analysis of implementation was broader than the conduct and management of controlled experiments, it, nevertheless, appears to be particularly relevant on that broader level. Most significantly, Fixsen et al. (2005) identified that “core implementation components, organizational components, and influence factors interact to produce implementation outcomes” (p.66). Their model depicted the relationship between the core implementation components and the outer core of organizational and influencing factors, which enable and sustain them. However, they commented that the evidence and, therefore, their analysis focused mostly on the ‘outer’ ring components and factors. They found far less evidence for the strategies and processes to secure the support of individuals, departments and agencies to secure successful implementation. Yet, they also appear to accept – and their model emphasises - that the responsibility for the successful operation of the crucial “implementation components” must lie with the operational managers, which has been a strong theme within the Turning Point case study.

The model, which seeks to integrate the novice categories (Chapter 3) and the tactical (Chapter 4), operational and strategic factors (this Chapter) for the conduct and management of an experiment, emphasises two key interdependencies: the researcher responsibility for the tactical or mechanical components of the trial; the iterative relationship between the research station, research motivation and the operational and strategic components. As Sherman (2010) and Strang (2012) have emphasised, social and intellectual foundations not only have to be built, they must be maintained, sustained and developed as the experiment progresses.

This, in turn, suggests that “novice theory” can only provide a partial explanation for treatment integrity, which is also dependent on those foundations and the protective factors we have set out in Table 5.4. To discuss this, further, we return to the three novice categories, Boruch, Sherman and the seven themes.

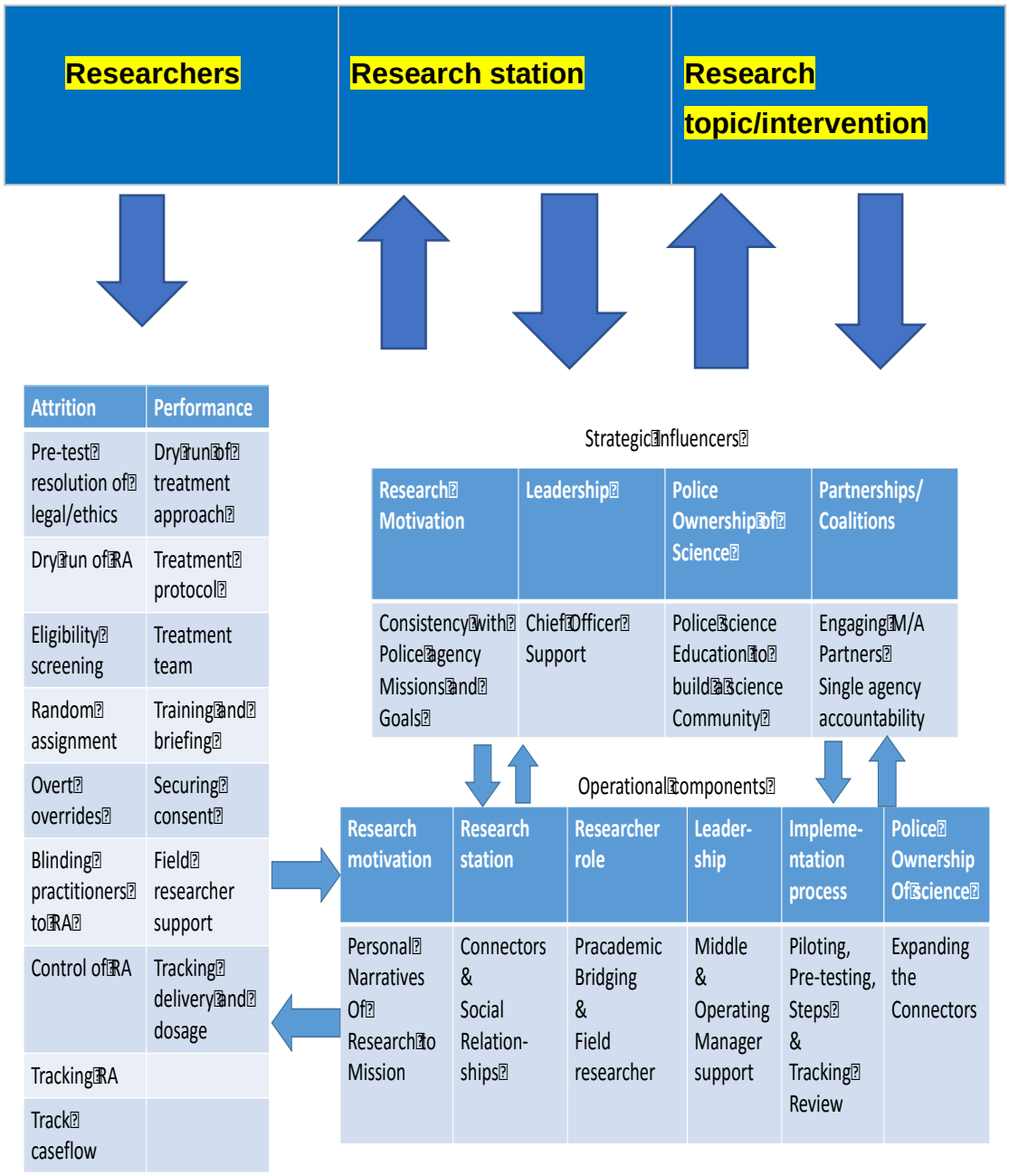


Figure 5.2: The novice, implementation components and narrative themes from Turning Point mapped against Fixsen et al. (2005) NIRN model of implementation

5.6.4. The research topic: research motivation and implementation.

We have suggested that Boruch implied an 'outsider' model of research, in which the research topic was proposed to the organisation rather than being generated from within. As a result, he placed considerable emphasis on the processes for engaging the organisation and incentivizing the institutional and strategic levels. Sherman (2010) stressed the importance of establishing an effective "contract" with the organisation at the outset, but also placed far more emphasis on the operational managers and staff. The interviews with the Turning Point participants stress the importance of paying attention to three levels: the institutional level, including the governance of the police agency; operational middle management; the frontline supervisors and staff and their engagement in the research motivation and implementation.

Whether the programme is initiated by an external sponsor (Boruch, 1997), created in partnership (Sherman, 2010) or by an internal "connector-innovator", Rogers (1995) and Sherman (2010) have identified the importance of the relevance and benefits of the research question. Rogers' diffusion theory (Rogers, 1995) proposed that certain characteristics – relative advantage, compatibility, complexity, trialability and observability – would influence the likelihood of and speed of adoption of an innovation. Focusing particularly on "relative advantage", Sherman (2010) distinguished "great" from "routine" science in an experiment by the extent to which a "given contribution" could make "a leap forward in understanding (with theory), intervention (with public benefits) or both" (p.401).

Sherman and Neyroud (2012) framed the research questions for Turning Point around the context – fiscal constraint – and the evidence that pointed to a potentially beneficial way of focusing police resources by targeting high

harm offenders on the one hand and diverting and encouraging desistance by low harm offenders on the other. The interviewees responses suggest that they understood both the relative advantage and the compatibility of the research proposal with the force's strategy and the austerity context in which they found themselves. As the experiment progressed, the context became progressively more advantageous as the Ministry of Justice launched first a consultation about a revised approach to OOCD's and then the three-force pilot (Ames et al., forthcoming), which were both influenced by Turning Point's design and early findings. As a result, even ahead of the formal evaluation, Turning Point had become a vehicle for change and West Midlands identified as an organisational innovator.

The active support of the Chief Officers was not only important in initiating and sustaining the research, but provided a mandate to middle managers and frontline staff engaged in the experiment which some actively brandished at reluctant colleagues to secure support and compliance. It is important to note that the relationship between the Chief Officers and the researchers was qualitatively different from the more normal insider to outsider relationship. The Chief Constable had been the author's deputy, when the author was the Chief Executive of the NPIA. One of the other key members of the Chief Officer group was a recent PEP student with a strong relationship with Cambridge. Whilst formal conversations and agreements were still important, the level of personal trust that existed before the research started undoubtedly eased the process of negotiating the experiment and overcoming difficult moments such as the restart at the beginning of Stage 4.

The importance of the institutional context goes beyond the Chief Officer team and the organisation to its governance and wider political context. Turning Point was started in the hiatus between two police governance systems. The Coalition government had announced the abolition of Police Authorities and their replacement by elected Police and Crime Commissioners (PCCs) in 2010 (Davies, 2014). The PCCs were elected in November 2012. The new PCC for West Midlands was from an opposition party to the Coalition government. The latter had ceded local control to the PCC as part of the

reform, leaving the PCC substantial freedom to shape local policing policy. The PCC was briefed about Turning Point before election and, in more detail, afterwards, once in office.

Crank and Langworthy (1992) have argued, in applying institutional theory to policing, that police departments are both "constrained and enabled by their institutional environment" (p.361). They suggest that, rather than engaging their "institutional sovereigns" with technical arguments about their efficiency and effectiveness, successful departments deploy powerful myths designed to connect with the external narratives about crime-fighting. Their observations were very much based on the US context of smaller, largely municipal police departments. However, the UK context, with the change from Police Authorities to PCCs, had shifted to a more US model with policy control devolved locally (Davies, 2014).

When the Chief Officer team and the researcher presented Turning Point to the PCC for the first time, it was as a technical, scientific experiment. However, the Chief Constable also framed it within a narrative which emphasised West Midlands as an innovative force committed to evidence-based approaches. At the first joint strategy meeting hosted by the new PCC and Chief Constable for senior staff in the force and at a subsequent SEBP event open to all staff, the Chief Constable explicitly deployed Turning Point as an example of the force's approach to using evidence as a way of finding more effective and efficient ways of delivering policing at a time of financial constraint. Consistent with Crank and Langworthy's emphasis on institutional mythology, the Chief Constable connected the scientific enterprise with meeting the challenge of austerity through an evidence-based approach.

As the experiment developed, the PCC's focus on the impact on victims became an important factor in driving the second RCT comparing victim experience in the treatment and control samples. For the PCC and his staff, who had been given responsibility for commissioning victim support services, the ability to present Turning Point as a positive gain for victims became important.

Research motivation needs, therefore, to be situated in the institutional, political and organisational context of the experiment. This alone suggests why a novice investigator working with unfamiliar and novice agency may encounter greater obstacles than a more experienced investigator in a department already committed to experimental research and prepared, as West Midlands was, to frame it within their overall strategic approach. However, this also suggests an area in which the "pracademic" has an advantage from being an "insider" with an insider's knowledge of such issues. The "pracademic" may also, as was the case with Turning Point, be a "connector" who has a network, built whilst undertaking post-graduate studies, connecting them to experienced researchers.

5.6.5. The researcher: the researcher role and "pracademics"

Turning Point, Checkpoint and the East Midlands experiment highlight a key difference in the institutional relationships where the research is driven from within the force by practitioner-led research. Boruch's model presumed that a "credible external research body", funded by a sponsor such as the National Institute of Justice, would seek to persuade a research site to let the researchers undertake the research. In Turning Point and Checkpoint, the organisational motivation for the research started in the middle. In Turning Point, it was an influential, middle ranking "connector"- "innovator" who started the process of initiation by linking the ideas (Sherman and Neyroud, 2012) with the institutional hierarchy.

The role of "connectors" was suggested by Gladwell (2002) as one of three roles critical to creating a "tipping point" in innovation or change. "Connectors" were defined by him by their knowledge of a network of people and ability to link people with an idea to those who might progress it. Rogers (1995 and 2002) proposed the importance of a small group of "innovators" as the initiators in adopting new practices. As the Chief Executive of the National Policing Improvement Agency from 2006 to 2010, the author had pursued a

deliberate policy of funding Cambridge PEP places in partnership with Sherman to seed the police service with “innovators” who would be able to act as the connectors between researchers and research opportunities.

A complementary strategy was adopted in Australia where the University of Queensland and Australian Institute of Police Management ran three annual evidence-based Masterclasses for senior and middle ranking representatives from Australia and New Zealand. Several of the attendees were then sponsored to attend the Cambridge PEP. Turning Point, Checkpoint and a growing number of the in-flight experiments in the UK and Australia would be appear to have been influenced, if not directly initiated, because of these strategies (Appendix 2).

The importance of such connector-researcher networks is highlighted by the problems with early implementation in the MoJ pilots (Ames et al., forthcoming). As we have seen in analysing “novice theory”, knowing how to manage an experiment effectively in the field is an important foundation of a high integrity experiment. By contrast, the MoJ pilots involved three novice forces and a research team new to field experimentation in policing. The evaluation report provides ample evidence that lessons from police field trials were not incorporated into the design or implementation of the pilots.

In contrast, West Midlands was a not a novice research station. However, their first experiment, a test of a targeted intervention to reduce repeat anti-social behaviour (see Appendix 2), had suffered from significant problems of treatment integrity in one of the two sites. On the other hand, this meant that, although, as one of the interviewees observed, the force knew that RCTs were very hard to implement, the middle managers in the first two LPAs to implement Turning Point had considerable understanding of the process and steps required. On that basis, the need for step by step piloting, the Cambridge “kick-off” event, the project structure and a system of tracking delivery were readily accepted.

The experience in the West Midlands has important parallels in the

experience in the Queensland Police Service. The first Queensland experiment was the QCET Legitimacy experiment (Mazerolle et al., 2012). That experiment was run largely to an 'outsider' model, with the research team from the University of Queensland. However, there was a critical senior 'connector' and 'pracademic', Assistant (now Deputy) Commissioner Dr Peter Martin, who helped sustain the experiment through some tough internal debates with Unions and middle managers. The most recent Queensland RCTs testing training (Platz and Thomson – Appendix 2), forensics (Antrobus and Pilotto, 2016), field technology and Body Worn Cameras (see Appendix 2) have used 'practitioners' and an embedded criminologist to conduct and manage the RCTs. Whilst the participants in the original RCT were largely unaware that they were part of a RCT (Bedford, personal communication), the succession of trials since have been sponsored by Chief Officers, have all involved practitioners within the research team and come against a backdrop of a now well established Australia and New Zealand SEBP Chaired by Deputy Commissioner Martin.

5.6.6. From the research station to building a "field station": leadership, partnerships, coalitions and the ownership of science

In describing the "novice" categories, we deliberately used the term "research station" to differentiate an agency hosting a single RCT from Sherman's vision of the "field station", which would provide the platform for a programme of experimental research. The findings from this (summarised in Table 5.4) and the earlier chapters suggest several important features of the "field station" which can be identified from the experience of Turning Point.

Sherman (2010) argued that the social foundations upon which the successful research programmes relied centred around four layers in the organisation ranging from the executive, through operational management to the treatment and case processing staff. Of these, we have already suggested the support of the executive was a critical factor in establishing and supporting the experiment. Whilst Boruch emphasised this level, Sherman (2010) drew on

his experience in Milwaukee (Sherman et al., 1992 and Sherman, 1992) to focus on the middle managers engaged in “operational liaison”. For Sherman, this was a role that needed to be established for the duration of the experiment. The best practice that he cited was the role that Lieutenant Collins played in Milwaukee in making the case for random assignment. In Turning Point, the research-practice boundary was more porous than Sherman’s description of Milwaukee. The presence of pracademics on both research and practice sides meant that advocacy for experimental approaches was shared.

Sherman’s Milwaukee case study (Sherman, 1992) also focused on another part of the middle management operational liaison role – the leadership and management approach required to conduct and manage a successful experiment. Sherman stated that the “lesson for future experiments is to seek host agencies with strong local leadership below the top executive level” (1992:317). These leaders needed to “strong and charismatic” and provide “motivation and inspiration”. Conversely, the literature on police reform suggests that middle managers have frequently proved a major obstacle (Greene, 2007, Skogan, 2008, and Marks and Sklansky, 2012). This research suggests the importance of “unfreezing” and warming up the “frozen middle”, who tend to be focused on transactional performance challenges (Byrne, 2005, Loftus, 2009 and Cockcroft, 2013), to the ideas of experimental criminology, its connection with the agency’s strategy and their central contribution to its success.

In his more recent work, Sherman has developed his thinking on the critical importance of the operational and middle management role in tracking compliance, feeding back and using focused performance management approaches such as “copstat” to ensure tight delivery of treatments (Sherman, 2013 and Sherman et al., 2015). Weisburd (2005) had earlier described the importance of Gajewski’s role in ensuring compliance in a hotpots study (Weisburd and Green, 1995). In a similar vein, Wood et al. (2014) and Sorg et al. (2016), analysing the experience of the two Philadelphia patrol experiments (Ratcliffe et al., 2011 and Groff et al., 2015) identified that their

problems with delivering dosage were strongly linked to middle managers – Captains - who, despite ordering officers to patrol their designated hotspots, nevertheless knowingly allowed them to stray to other areas.

The analysis of the Turning Point interviews and the case study in Chapter 4 suggest that inspirational leadership, tracking and ownership are all important contributors to effective conduct and management of a police RCT. The most important of the implementation components that we identified in Chapter 4 as contributing to treatment integrity appear to have been those which supported an active process of tracking eligibility screening, random assignments and treatment delivery, feeding back (through individual feedback, briefing or training) and following through to ensure that behaviour change has occurred. The research team (Slothower et al., 2015) have described how important this cycle of "TTF" (Tracking, Training and Feedback) was in "improving compliance with policy and improving policy itself" (p. 109). This also appears to find support from Grimshaw et al. (2012), who systematically reviewed the most effective strategies for translating research into clinical practice in medicine. They demonstrated that combined strategies that included audit and feedback, checklists and reminders and generating a "pull" appetite for research findings to use in practice, were linked to higher levels of translation and implementation.

It is not just the management mechanisms that appear to matter but also the way the tracking process was performed by the Custody Inspectors and OMT Sergeants. The Custody Sergeants and OMT members appeared to find the process supportive rather than threatening. Both groups, in their interviews, seemed quite happy that their decisions through the Gateway exclusions and Turning Point treatment plans were being scrutinized and they were getting phone calls or emails asking them to explain or giving them feedback. Indeed, two OMT interviewees specifically mentioned that a major benefit of the "prescribing tool" had been the visibility of their Turning Point contracts to the supervisors and the support provided as a result.

As we saw in Chapter 4, despite the serious attrition problems in Stage 3, the

research and project team made strenuous efforts not to blame the custody officers and to concentrate on learning rather than sanctions. This was, in part, a product of the Chief Constable's very overt commitment to his staff that if they used their discretion, they would be supported. It was also consistent with the growing body of research on the importance of procedural justice within the organisation as well in operational practice. Wolfe and Piquero (2011), Bradford and Quinton (2014) and Haas et al. (2015) have shown, within US, UK and Brazilian police departments, that levels of compliance with managerially set requirements and departmental rules are associated with the officers' perceptions of the fairness of the management ethos in the agency.

This emphasis on fairness and transparency has clear implications for the leadership styles in RCTs. Strang (2012) suggested that leaders involved with RCT "coalitions" needed to be able to deploy a combination of transactional and transformational leadership approaches in order to achieve both tight implementation and provide a clear vision of the connection between the research and the agency's strategy. The legitimacy research and the Turning Point case study and interviews both add weight to Strang's argument. But it also further suggests a need to incorporate lessons from "authentic" models of leadership, which emphasise a greater focus on moral reasoning and relational transparency than transactional and transformational models (Walumba et al., 2008, Neyroud, 2011 and Iszatt-White and Saunders, 2014). Consistent with this, a notable feature of the leadership approach in Turning Point, which was valued by many the interviewees, was the time invested in explanation and the time allowed to help "sceptics cross the bridge" towards support for the experimental approach.

Indeed, this study suggests that the effectiveness of the experiment was linked to much more than compliance. Recent analyses of other experiments have reached similar conclusions. Famega et al. (2016) suggested that the officers' behaviour in complying with the required dosage of patrol in their hotspot study was linked to their ownership of the science and design of their contribution to the experiment. Sorg et al. (2016) supported this and stated that "input of line officers into the design of experiments could help to

anticipate and mitigate officer non-compliance with treatment conditions" (p. 22). Conversely, the RAND process evaluation of a National College of Policing RCT of procedural justice training for Stop and Search (Giacomantonio et al., 2016) found that shortfalls in implementation appeared to be linked to the failure to explain the purpose of the trial or connect it to officer's own sense of mission.

The sense of ownership in Turning Point seemed to have had two connected levels: the ownership of key design features of the experiment; ownership and valuing of the scientific enterprise of the experiment. The first appears to have been founded in the steps that the research and project team took to engage Custody Staff and OMT's in the design of the Gateway, the exclusion criteria, the treatment approach and the "prescribing tool". The second appears to have started from the investment in an intensive "kick off" – the Cambridge sessions. The RAND team noted that the absence of such a collective ritual in the Stop and Search study (Giacomantonio et al., 2016). They found that the reliance on the individual efforts of local liaison Inspectors had diminished the impact of the experiment. In Turning Point, the "kick-off" was gradually reinforced by the series of training events, personal conversations with the research team and the larger SEBP-led events which developed as the experiment progressed. Rynes (2007) has argued that the latter – practitioner-researcher seminars supported by a professionally led body (in this case the SEBP) – can support the "tipping point" towards evidence-based practice, broadening the small group of innovators into a larger group of potential supporters.

5.7. Conclusion: From "protective factors" to "Field Universities" of Policing?

This chapter has explored two research questions: the extent to which it is possible to construct a model of the key protective factors that contribute to high levels of treatment integrity in police RCTs; the extent to which those factors add to or modify "novice theory". From the coding and analysis of the

interviews with a sample of key staff involved in the case study RCT – Operation Turning Point – we have constructed a model of the protective factors (summarised in figure 5.2). We have also, in response to the second question, suggested that novice theory needs to be understood in the context of both the operational and protective factors that we have identified. Taken together these findings indicate the potential advantages of building institutional frameworks in which the development of practitioners and researchers and the conduct and management of experimental research could be brought closer together.

Sherman has proposed "field stations", Shepherd (2003) a police station version of the teaching hospital and Weisburd and Neyroud (2011) highlighted the importance of the police taking ownership of science. The model which has emerged from the analysis of 122 police RCTs, in general, and Operation Turning Point, in particular, appears to suggest the advantages of developing "field universities" of policing, which combine all three approaches: an active and committed "field station" supported by the leadership, management, social and intellectual foundations that we have described above; a teaching and research environment in which "pracademics" and "aca-titioners"[8] provide teaching to develop police science understanding and research to develop police science knowledge; a new profession of policing committed to scientific curiosity and challenging and testing practice (Sherman and Neyroud, 2013).

[8] The term "aca-titioner" was used by Dr Justin Ready at the 2017 Society for Evidence Based Policing Conference to describe academics who have a strong practitioner background or even a current practice commitment.

6. Conclusions and Recommendations

6.1 Conclusions:

This research has been focused on randomised controlled trials involving the police. Although the first police RCTs were completed nearly fifty years ago in the late 1960's and early 1970's (Rose and Hamilton, 1970, Earle, 1973 and Stratton, 1975), the real expansion in numbers of completed RCTs has occurred very recently, since 2010. RCTs have been controversial in other fields (Deaton and Cartwright, 2016), but even the relatively small numbers of RCTs in policing have generated considerable controversy (Manning, 2005, Cockbain and Knuttson, 2014 and Greene, 2014). Much of the critical reflection on RCTs has concentrated on their general validity. The arguments have tended to centre on the combination of the complexity of the policing context and the controlled manipulation required in a RCT, which have cast doubt on the ability to replicate findings in wider practice (Hope, 2009, Sampson et al., 2013 and Laycock and Mallender, 2015). However, amidst the first wave of RCTs in policing and criminal justice, Clarke and Cornish (1972) and Martinson (1974) had drawn attention to a potentially much more serious flaw – the challenges of securing the internal validity of the experiment.

We have argued, drawing on Sherman (2010), that the primary test of a successful experiment should be the extent to which the trial was delivered with high levels of treatment integrity: in particular, low attrition from the randomised sample and high performance in delivering the intended treatments. Eck (2002) argued that the extent to which this was successfully achieved would always be a trade-off between the need to maintain "real world" conditions to support greater external validity and high levels of control to ensure internal validity: a trade-off he christened the "diabolical dilemma". This analysis has suggested that Eck may have overstated this case. Instead, we have argued that the solution to the dilemma lies in building a better understanding of the methods to conduct and manage police RCTs with high levels of treatment under field conditions and, in doing so, provide

practitioners with better models for delivery of the proposed intervention. We suggest that there is no simple dichotomy between experimental research and the implementation of tested practice: a successful experiment should not just be able to test the intervention, but also to shape the wider implementation approach.

The study of police RCTs has become all the timelier because of two factors: the growing importance of an “evidence-based policing” movement in policing; the increase in the number of practitioners and new researchers in conducting RCTs in policing. Up until 2010 there had only been four RCTs published in which practitioners featured on the list of authors (Earle, 1973, Sherman et al., 1992, Braga et al., 1999 and Shipley and Baranski, 2002). Braga et al. (2014) also showed that the network of scholars responsible for completing police RCTs had been a small one, centred around an even smaller group of expert experimentalists. Since 2010 the network of scholars has expanded and practitioner involvement has grown to the point that two thirds of the current ‘in-flight’ police RCTs are being led or co-investigated by practitioners (Appendix 2). A major factor contributing to this has been the significant emphasis on and increase in the level of post-graduate education in policing over the last decade (Neyroud, 2011 and Konze, 2016).

Despite this and, unlike the field of medical experimentation, there is relatively little systematic knowledge of the completed police RCTs or the best methods to conduct and manage them successfully. First and foremost, there is no national or international register of police experiments, nor is there a single recognised textbook on conducting and managing RCTs in policing. Whilst there are a number of important methods works on experimental research (for example, Shadish et al., 2002, Torgerson and Torgerson, 2008 and Glennerster and Takavarasha, 2013), they all tend to focus on the research design and statistical methods rather than the conduct and management of the experiment. There is a much smaller body of work which addresses the challenges of field experimentation in criminal justice and policing and the methods to secure high levels of treatment integrity: of these Dennis, 1988 and 1990, Boruch, 1997 and Sherman, 2010 stand out. Sherman (2010), in

particular, has provided the most practical and grounded approach, drawing on his experience as one of Braga et al. (2014)'s expert experimentalists.

Sherman's grounded approach has, in turn, helped to shape this research. However, it was also motivated by the author's own personal experience and, initially, failures, as a novice experimentalist in the field managing Operation Turning Point. Despite more than thirty years as a police officer – for more than a decade of which the author was a Chief Officer – field experimentation in policing proved more complex and challenging than many of the major national programmes for which the author had been responsible. As a result, the experience of managing the field research changed the focus of the author's own research from the specific evaluation of one RCT to the analysis of the wider field of police RCTs and their effective conduct and management.

The research has, therefore, been designed to fill key gaps in our knowledge about police RCTs - in particular the numbers, topics and treatment integrity of completed RCTs – and build a more comprehensive, grounded theory of the key factors which contribute to the successful conduct and management of police RCTs. Following these themes, the research has been presented in four chapters which have been supported by three different methods – systematic search and meta-analysis of the attrition and treatment performance, case study and grounded theory – to explore a set of connected research questions on the conduct and management of successful police RCTs with high treatment integrity.

In Chapter 2 and Appendices 1 and 2, drawing on a search for randomised trials in policing, we have provided the first comprehensive dataset of police involved RCTs – both those completed (by 2016) and those that have been "in flight" during the research timeframe (which includes those completed but not published before the start of 2016). The completed dataset has been presented in a "matrix" (Appendix 1) which summarises each RCT and provides the data extracted from each report or article about the level of treatment integrity in the experiment. For a significant proportion of the studies we were not able to make a reliable assessment, because the authors have

provided either insufficient or unclear data. However, we have been able to provide treatment integrity ratings for more than two thirds of the 122 police RCTs that we found in the search.

The analysis in Chapter 2 has provided a number of significant findings. Some of these build on the findings from Braga et al. (2014) who were able to show an accelerating growth in police RCT numbers and a range of topics. We have been able to show that growth has continued to accelerate and that the range of topics was wider than and has expanded beyond Braga et al. (2014)'s initial list. By focusing on treatment integrity, we have also been able to show a matching increase in treatment integrity until the more recent rapid growth in numbers.

From this analysis, we drew out a set of data that suggested that the flattening off in treatment integrity may be the result of a combination of three factors that we found to be associated with low treatment integrity: novice researchers; new police agencies becoming research stations; novel interventions being tested. In Chapter 3 we tested this theory by analysing the extent to which "novice theory" was predictive of high or low treatment integrity. The analysis suggested that whilst the novice categories provided an explanation for the overall patterns of treatment integrity, a more detailed analysis comparing some of the topic areas suggested that other factors were also important.

We were also able to identify a group of three topics from the 122 RCTs which seemed to have particular problems with treatment integrity. For two of these – hotspots and Body Worn Video studies – the reasons and mitigation were apparent. With the notable exception of Sherman and Weisburd's (1995) original Minneapolis study, there had been insufficient attention paid to the level of patrol treatment applied to each hotspot. More recent studies, using techniques such as GPS monitoring (Ariel et al., 2016), have been able to provide much more reliable estimates. In the case of Body Worn Video, none of the four early studies (Ariel et al., 2014, Owens et al., 2014, Grossmith et al., 2015 and Jennings et al., 2015) were able to provide reliable estimates of

the extent to which the equipment had been deployed. The most recent studies have placed a much greater emphasis on tracking deployment (Drover and Ariel, 2015 and Ariel et al., 2016).

The third topic, juvenile justice diversion studies, showed consistently lower treatment integrity levels than other topics. The diversion studies all used a different design – a trickle flow of cases – to the hotspot and Body Worn experiments. Moreover, it was quite apparent from the analysis that there were a number of common features, which were also present in the Turning Point experiment that the author had been conducting. The diversion studies had problems with both attrition and treatment delivery. A number of factors appeared to contribute to this: the case flow depended on a consistent exercise of officers' discretion in screening cases in to the experiment; most of the experiments reported evidence of officers reassigning cases and covertly overriding the experimental protocol; treatments often had to be tailored to the individual offender and there was substantial evidence of inconsistent delivery.

In Chapter 4, we used a detailed analysis of the low integrity diversion studies, together with a review of a number of published case studies and commentaries on the conduct and management of police RCTs, to provide a framework for a case study on Operation Turning Point. The latter experiment provided an important opportunity to explore the factors that contribute to high integrity. The experiment had, in its first intended evaluation phase, suffered from low levels of treatment integrity due to both high levels of attrition and problems with consistent treatment delivery. In contrast, after the experiment was restarted and changes were implemented, Stage 4 saw high levels of treatment integrity.

The case study focused on the "implementation components" which appeared to have contributed to this change. We found that the steps taken in the first three Stages (as set out in Tables 4.1 and 4.4) were necessary but not sufficient to ensure a high level of treatment integrity: resolving the ethical and legal issues was essential to initiate the experiment: piloting and pre-testing

the eligibility screening and randomisation Gateway was absolutely necessary to learn lessons and establish the Turning Point treatment model; training, in field support, tracking of eligibility overrides, a system of email alerts and monthly meetings provided the basis for a model of training, tracking, feedback, support and monitoring. Comprehensive though these appeared to be, they did not prevent high levels of attrition and inconsistency in treatment delivery.

However, the significant difference in treatment integrity between the Consort for Stage 3 (Figure 4.1) and Stage 4 (Figure 4.2) suggested that we needed to focus on the changes in conduct and management of the experiment which were instituted when the experiment was restarted in Stage 4. These suggested that the combination of addressing the technology support, revising the Gateway, providing active field support and more active tracking of cases appears to have redressed the shortcomings in eligibility screening and random assignment sufficiently to achieve a low level of attrition. Equally, the development and deployment of a “prescribing tool” appears to have transformed the level of treatment delivery.

The first part of the case study focused on the detailed mechanisms contributing to high levels of treatment integrity. It also showed, as Fixsen et al. (2005) found in reviewing implementation of social programmes, that there were wider operational and strategic influences that were also important to initiate and sustain a successful experiment. In particular, it was apparent that the level of senior and middle management support combined with the frontline commitment and ownership of the experiment had been a significant factor. Without ownership, for example, there would have been no “prescribing tool” which had been developed by the Offender Manager teams themselves to overcome the researchers’ observations about the problems with their treatment consistency. Famega et al. (2016) had also found that the ownership of the experiment and its science was an important contributor to treatment delivery levels in a hotspot experiment.

As the Turning Point experiment developed, West Midlands Police became progressively more interested in understanding both the outcomes of the experiment and the lessons that could be learnt from the implementation. As a result of this, the research team were given approval to interview a sample of those involved in the experiment. The author was able to design the interview process to explore the police officers' perceptions of a number of areas identified in prior studies of the operational implementation of police RCTs (notably Boruch, 1997, Sherman 2010 and Strang, 2012). The interviews were coded and the initial categories developed into a set of narrative themes drawn from the interviewees' observations.

Together with the findings of Chapters 2 and 3 and the model of implementation mechanisms developed in Chapter 4, these themes have been used to construct a model of the wider protective factors necessary for conducting and managing effective police RCTs, which was set out in the second half of Chapter 5. The model situates the "novice theory" which we developed from the analysis of the 122 RCTs in a framework combining experimental mechanisms with operational and strategic factors (Figure 5.2).

There are some limitations to the model and its external validity which we have discussed in detail in Chapter 5. Above all, we need to acknowledge that Operation Turning Point was an unusually complex and extended experiment. It had to be restarted after the problems in Stage 3. It used a computer based Gateway to screen eligibility and randomise and relied on a large group of more than 90 police custody officers to do this. Its trickle flow design and pre-court diversion intervention would also mean that some aspects of the model may not be directly transferable to batch and cluster randomised designs or in dealing with other types of intervention. However, the lessons about the treatment integrity risks from testing new interventions suggests that the step by step approach to implementation which was adopted in Turning Point might be particularly important in such cases.

Alongside the questions of the external validity of the research, there are also potential limitations and possible bias as a result of the status of the

researcher. There are very few Chief Officers who have conducted and managed a police RCT in the last fifty years. The only others were Earle (1973) who was the deputy sheriff in LA County and Farrar who was Chief of Rialto at the time of the first Body Worn Video trial in that department (Ariel et al., 2014). Both were able to use their dual role to shape the research motivation and implementation whilst also acting as a principal investigator. Although this author had formally retired before the start of the experiment, the interviewees made it very plain that they regarded the author as a "pracademic", Chief Constable and researcher, bridging the role of senior practitioner and researcher. Yin (2014), in discussing the qualities necessary for case study research, suggested that there is a tight path to be trod between the importance of field and subject expertise and the ability to be an independent observer. In order to keep that balance, it was important in this research that the interviews were conducted independently, because it is difficult to see how the author's status would not have biased the responses. Moreover, throughout the research, the author has sought to triangulate observations with findings from the interviews, literature or documentary evidence from Turning Point. However, these research findings would be significantly strengthened by further research on other police RCTs.

6.2 Recommendations

The recommendations set out below are intended to provide a top ten list of the key factors distilled from this research that researchers and police agencies need to pay attention to in order to be able to conduct effective experiments in the future.

6.2.1 Publish a Crimport:

The discipline and transparency of a CRIMPORT protocol is an essential part of the scientific and operational process. As we have shown with Turning Point, the CRIMPORT provided clarity from the outset about the purpose, method, treatment standards and measurement of the RCT.

6.2.2 Provide training and support for the Principle Investigators

In Chapter 3 we demonstrated, as Feder et al. (2000) found, that novice investigators present a higher risk of low treatment integrity RCTs. The Campbell Collaboration (of which the author is the Co-Chair for the Crime and Justice Coordinating Group) encourage all authors registering titles to undertake approved training in meta-analysis and in the Campbell standards. The ASC Divisions of Policing and Experimental Criminology and the Society for Evidence should consider setting up a similar opportunity for experimental researchers. Adding to this with mentoring support from experienced RCT researchers would help to ensure that more novice researchers were successful in their first experiment. It would also help to grow the network that Braga et al. (2014) have shown to be so critical to experimental criminology.

6.2.3 Build Field Stations

Whilst novice researchers are the clearest risk to treatment integrity, establishing experiments in agencies which have no prior experience of RCTs is a further risk. We have shown the benefits, in the Turning Point experiment, to building a field station within an agency in which a programme of experiments are conducted. At the time of writing, West Midlands Police has embarked on a further four field experiments, using largely internal expertise but with support from this researcher. The systems, processes and understanding of the experimental discipline is now well enough embedded for such practitioner-led research to be a realistic proposition.

6.2.4 Test new interventions in stages

The third major risk factor that we identified in Chapter 3 was the testing of novel interventions. It is clear from the analysis of Turning Point and the experience of research in new topic areas such as Body Worn Video that a

staged process of testing is most likely to secure high levels of treatment integrity: a small-scale pre-test of the viability and operation of the intervention; an efficacy test under highly controlled conditions; effectiveness testing under operational conditions and preferably in more than one agency.

6.2.5 Test AND Track Eligibility screening and Randomisation

In Chapter 4 we identified that the design and management of the eligibility screening and random assignment processes were critical to successful experiments. The combined eligibility screening and random assignment tool in Turning Point was tested in stages. This was necessary but not sufficient to prevent the problems encountered in Stage 3. It was only the active tracking of cases, combined with the redesigned tool, that ensured high levels of treatment integrity in the final Stage.

6.2.6 Build in overt overrides

As Dennis (1988) showed, designing mechanisms for overt overrides on eligibility and treatment delivery is an essential means of combining professional discretion with the control of the experiment. The more that the frontline staff involved in the experiment are engaged in the design and development of the overrides the better. Furthermore, drawing on Hobday's (2015) research, the feedback from the overrides can be an important part of building confidence and understanding of the experiment.

6.2.7 Train, Track and Feedback

Sherman, Slothower and Neyroud (2015) have shown the benefits of a disciplined and purposive cycle of training, tracking and feedback. In Turning Point this process was repeated several times with an initial training event being followed by further training at each Stage as lessons were learnt and

the processes revised. The training went well beyond drilling staff in the details of the intervention. It was deliberately designed to build an understanding of the experimental science.

6.2.8 Build Ownership of the Science

Weisburd and Neyroud (2011) and Famega et al. (2016) argued that police ownership of the science of policing is essential to both the effective conduct of experiments and the reform of policing. The Turning Point case study suggested that ownership of the science, both by leaders and frontline staff, was a key protective factor that supported high levels of treatment integrity.

6.2.9 Use "Pracademics" to conduct the research

In Chapter 2 we showed how there has been a significant shift over the last decade to practitioner led experimental research. The research presented in this thesis has demonstrated the benefits of "pracademic" research in a field which had, until quite recently, relied almost exclusively on "outsider" research.

6.2.10 Create new "Field Universities of Policing"

Finally, if we bring several these recommendations together – the need for field stations, the benefits of programmes of testing, the importance of the police ownership of science and the engagement of practitioners in the process of research and testing – it is apparent that the future of police RCTs would benefit from the development of "field universities" of policing. In our recent article (Bedford and Neyroud, 2017) we have shown how conducting field experiments in policing can build organisational learning and contribute to the improvement of practice. The "field university" which was first

suggested by Shepherd (2003) offers the opportunity for policing, as in medicine, to bring learning, testing and operational practice closer together to their mutual benefit.

The analysis presented in this thesis has suggested that experimental police science may be at an important tipping point, in which the volume of police RCTs and the body of experienced researchers has created the opportunity to test and replicate the testing of police practice on a much wider scale than previously possible. Whether that happens, as it has done in medical science, will, in our view, depend on the development of clusters of “field universities” of policing, in which programmes of research, teaching and scientific discussion of policing become normal business rather than individual experiments in isolated departments. This study suggests that policing may have moved closer to Shepherd’s (2003) vision of field science driving practice, but that there is still some way to travel.

7. Bibliography

7.1 General Bibliography

ACPO (Association of Chief Police Officers). (2009) *Gravity Factors Matrix for Out of Court Disposals*, London: ACPO.

Akobeng, A.K. (2005) 'Understanding Randomised Controlled Trials', *Archives of Disease in Childhood*, 90: 840-844.

Alpert, G. P., Rojek, J. and Hansen, J. A. (2013) *Building Bridges Between Police Researchers and Practitioners: Agents of Change in a Complex World*, Washington, DC: National Institute of Justice.

Ames, A., Di Antonio, E., Hitchcock, J., Webster, S., Wong, K., Ellingworth, D., Linda Meadows, L., McAlonan, D., Uhrig, N. and Logue, C. (forthcoming) *Adult Out of Court Disposal Pilot Evaluation: Final Report*, London: Ministry of Justice.

Andrews, D. A. and Bonta, J. (2010) *The Psychology of Criminal Conduct*. 5th Edition, New Providence, NJ: Lexis Nexis.

Angrist, J.D. (2006) 'Instrumental variables methods in experimental criminological research: what, why and how', *Journal of Experimental Criminology*, 2: 23-44.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association,* 91(434): 444-455.

Ariel, B. and Farrar, J. (2013) *The Effects of Body Worn Video on Police Citizen Encounters: An RCT*, [Presentation to the 2013 Conference on

Evidence-based policing, July 4th 2013], Retrieved 11th December 2016 from www.crim.ac.uk/events/ebp/2013.

Ariel B., Sherman, L. and Vila, J. (2012) 'Random Assignment without Tears: How to Stop Worrying and Love the Cambridge Randomizer', *Journal of Experimental Criminology,* 8(2): 193-208.

Barnes, G. C., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E. and Malvestuto, R. (2010) 'Low-intensity community supervision for low-risk offenders: a randomized, controlled trial', *Journal of Experimental Criminology,* 6: 159–189.

Bedford, L. and Mazerolle, L. (2014) 'Beyond the Evidence: Organizational Learning from RCTs in Policing', Policing, 8(4): 402-416.

Bedford, L. and Neyroud, P.W. (2017) Organisational learning from field research in policing: how police can improve policy and practice by implementing randomised controlled trials. *Police Practice and Research*, 19, 1-34.

Benner, P. (1982) From Novice to Expert. *American Journal of Nursing, March,* 402-7.

Berk, R.A. (2005) 'Randomized experiments as the bronze standard'. *Journal of Experimental Criminology,* 1: 417-433.

Berk, R.A. and Sherman, L.W. (1988) 'Police responses to Family Violence Incidents: An analysis of an experimental design with incomplete randomization' *Journal of the American Statistical Association*, 83(401): 70-76.

Berk, R., Sherman, L.W., Barnes, G., Ahlman, L., Kurtz, E. and Malvestuto, R. (2009) 'Forecasting Murder within a Population of Probationers and Parolees: a high stakes application of statistical learning', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172: 191–211.

Berk, R.A., Smyth, G.K. and Sherman, L.W. (1988) 'Why Random Assignment Fails: Some Lessons from the Minneapolis Spouse Abuse Experiment', *Journal of Quantitative Criminology,* 4(3): 209-223.

Blakeborough, L. and Pierpont, H. (2007) '*Conditional cautions: An examination of the early implementation of the scheme',* London: Ministry of Justice.

Boruch, R.F. (1997) *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Thousand Oaks, CA: Sage.

Boruch, R.F., McSweeney, A.J. and Soderstrom, E.J. (1978) 'Randomized field experiments for program planning, development and evaluation', *Evaluation Quarterly*, 2: 655-695.

Boydstun, J. (1975) '*The San Diego field interrogation experiment',* Washington, DC: Police Foundation.

Bradford, B. and Quinton, P. (2014) 'Self-legitimacy, Police Culture and Support for Democratic Policing in an English Constabulary', *British Journal of Criminology,* 54, 6: 1023-1046.

Bradley, D. and Nixon, C. (2009) 'Ending the 'dialogue of the deaf': evidence and policing policies and practices: an Australian case study', *Police Practice and Research: An International Journal,* 10(5-6): 423-435.

Braga, A.A. (2016) 'The Value of 'Pracademics' in Enhancing Crime Analysis in Police Departments', Policing, 10(3): 308-314.

Braga, A.A., Welsh, B.C., Papachristos, A.V., Schnell, C. and Grossman, L. (2014) 'The Growth of Randomized Experiments in Policing: The Vital Few and the Salience of Mentoring', *Journal of Experimental Criminology*, 10(1): 1-29.

Braucht, G. N. and Reichardt, C. S. (1993) 'A computerized approach to trickle-process, random assignment', *Evaluation Review,* 17(1): 79–90.

Brown, J. (1996) 'Police Research: Some Critical Issues', In Leishman, F., Loveday, B. and Savage, S. (Eds.) *Core issues in Policing*, London: Longman, pp. 177-190.

Byrne, J. (2005) '*Middle Management Excellence'*, Cambridge, MA: Harvard Business School. Retrieved 11th December 2016 from http://hbswk.hbs.edu/archive/5126.html.

Campbell, D.T. and Boruch, R.F. (1975) 'Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects', In Bennett, C.A. and Lumsdaine, A.A. (Eds.) *Evaluation and experience: Some critical issues in assessing social programs*. New York: Academic Press, 195-297.

Campbell, D.T. and Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research*, Chicago: Rand McNally.

Campbell Collaboration (2014) *Campbell Collaboration Systematic Reviews: Policies and Guidelines,* Retrieved 21st January 2017 from https://www.campbellcollaboration.org/images/library/campbell_policies_and_guidelines.pdf.

Canadian Council of Academies (2014) *Policing Canada in the 21st Century: New Policing for New Challenges. Ottawa (ON): The Expert Panel on the Future of Canadian Policing Models*, Ottawa: Council of Canadian Academies.

Cartwright, N. and Hardie, J. (2012) *Evidence-based policy: A practical guide to doing it better*, Oxford: Oxford University Press.

Carlier, I.V.E., Lamberts, R.D., Van Uchelen, J.J. and Gersons, B.P.R. (1998). 'Disaster-related post-traumatic stress in police officers: a field study of the impact of debriefing', *Stress Medicine*, 14: 143-148.

Carr, P.J. (2010) 'The problem with experimental criminology: A response to Sherman's 'Evidence and Liberty', *Criminology and Criminal Justice*, 10(1): 3-11.

Chan, J. (1997) *Changing Police Culture: Policing in a Multi-cultural society*, Cambridge: Cambridge University Press.

Chitty, C. (2005) 'The impact of corrections on reoffending: Conclusions and the way forward', In Harper, G. and Chitty, C. (Eds.) *The impact of corrections on re-offending: A Review of 'what works',* Home Office Research Study No. 291. London: Home Office.

CJJI (Criminal Justice Joint Inspectorate) (2011) *Exercising Discretion: The Gateway to Justice,* London: Criminal Justice Joint Inspectorates.

CJJI (Criminal Justice Joint Inspectorate) (2014) *A Joint Inspection of the Integrated Offender Management Approach*, London: Criminal Justice Joint Inspectorate.

Clarke, R.V.G. (2004) 'Technology, Criminology and Crime Science', *European Journal on Criminology, Policy and Research,* 10: 55-63.

Clarke, R.V.G. and Cornish, D.B. (1972) *The Controlled Trial in Institutional Research – paradigm or pitfall for penal evaluators?* Home Office Research Studies No. 14. London: Home Office.

Coalition for Evidence-based Policy (2015) *Top Tier Evidence Initiative: Identifying Social Programs Backed by Strong Evidence of Important Impacts on People's Lives*, Retrieved 2nd December 2016 from http://www.toptierevidence.org.

Cockbain, E. and Knutsson, J. (Eds.) (2014, *Applied Police Research: Challenges and Opportunities.* Abingdon: Routledge.

Cockcroft, T. (2013) *Police Culture: themes and concepts*, Abingdon: Routledge.

Cockcroft, T. and Beattie, I. (2009) 'Shifting cultures: managerialism and the rise of "performance"', *Policing: An International Journal of Police Strategies and Management,* 32(3): 526-540.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin.

Cornish, D.B. and Clarke, R.V.G. (1975) *Residential treatment and its effects on delinquency*, Home Office Research Studies No.32., London: HMSO.

Crank, J.P. and Langworthy, R. (1992) 'An Institutional Perspective of Policing', *Journal of Criminal Law and Criminology,* 83 (2): 338-363.

Cresswell, J.W. (1998) *Qualitative Inquiry and Research Design: Choosing among Five Traditions*, Thousand Oaks, CA.: Sage.

Criminal Justice Service (2004) *The Victim's Charter: a statement of service standards for victims of crime,* 2nd Edition, Retrieved 9th December 2016 from http://cgm-probation.org.uk/wp-content/uploads/2013/04/victims-charter2835.pdf.

Crown Prosecution Service (2016) *Conditional Cautioning Code of Practice*, London: Crown Prosecution Service. Retrieved 9th December 2016 from https://www.cps.gov.uk/publications/others/conditionalcautioning04.html.

Cullen, F.T. (2013) 'Rehabilitation: Beyond Nothing Works', In Tonry, M. (Ed.). *Crime and Justice in America 1975-2025*, Vol. 36 of Crime and Justice: a Review of Research, Chicago: University of Chicago Press.

Dall'Alba, G. and Sandberg, J. (2006) Unveiling Professional Development: A Critical Review of Stage Models. *Review of Educational Research, 76, 3*, 383-412.

Davies, M. (2014) *Unravelling the Role of Police and Crime Commissioners.* [Papers from the British Criminology Conference July 2014: 17–30].

Deaton, A. and Cartwright, N. (2016) *Understanding and Misunderstanding randomised controlled trials*, Cambridge, MA: National Bureau of Economic Research, NBER Working Paper No. 22595, Retrieved 9th December 2016 from http://www.nber.org/papers/w22595.

Dennis, M.L. (1988) 'Implementing Randomized Field experiments: an analysis of criminal and civil justice research', [Unpublished Ph.D. thesis, North-Western University].

Dennis, M.L. (1990) 'Assessing the validity of randomized field experiments: An example from Drug Abuse Treatment Research', *Evaluation Review,* 14(4): 347-373.

Disley, E., Taylor, C., Kruithof, K., Winpenny, E., Liddle, M., Sutherland, A., Lilford, R., Wright, S. and McAteer, S. (2016) *Evaluation of the Offender Liaison and Diversion Trial Schemes*, Santa Monica, CA: RAND. Retrieved 9th December 2016 from http://www.rand.org/t/rr1283.

Doran, G. T. (1981) ‘There's a S.M.A.R.T. way to write management's goals and objectives’, *Management Review*, 70(11): 35-36.

Dreyfus, S. and Dreyfus, H. (1980) *A Five-Stage Model of the mental Activities Involved in Directed Skills Acquisition*. Berkeley: University of California and the U.S. Air Force, Office of Scientific Research.

Drover, P. and Ariel, B. (2015) ‘Leading an experiment in police body-worn videos’, *International Criminal Justice Review,* 25(1): 80-97.

Dunford, F.W. (1990a) ‘Random Assignment: practical considerations from field experiments’, *Evaluation and Program Planning,* 13: 125-132.

Dunford, F.W. (2000) ‘Determining Program Success: the importance of Employing Experimental Research Designs’, *Crime and Delinquency,* 46(3): 425-434.

Durlauf, S. and Nagin, D. (2011) ‘Imprisonment and Crime: Can both be reduced?’, *Criminology and Public Policy,* 19(1): 13-54.

Durlak, J.A. and DuPre, E.P. (2008) ‘Implementation Matters: a Review of research on the influence of Implementation on Program Outcomes and Factors affecting Implementation’, *American Journal of Community Psychology*, 41: 327-350.

Eck, J.E. (2002) ‘Learning from experience in Problem-Oriented Policing and Situational Prevention: The positive functions of weak evaluations and the negative functions of strong ones’, *Crime Prevention Studies,* 14: 93-117.

Easton, H., Slivestri, M., Evans, K., Matthews, R. and Walklate, S. (2010) *Conditional Cautions: Evaluation of the women specific condition pilot,* London: Ministry of Justice Research Series 14/10.

Edwards, S.J.L., Lilford, R.J., Jackson, J.C., Hewison, J. and Thornton, J. (1998) 'Ethical Issues in the Design and Conduct of randomised controlled trials', *Health Technology Assessment,* 2(15): 1-146.

Eisner, M. (2009) 'No Effects in independent prevention trials: can we reject the cynical view?', *Journal of Experimental Criminology,* 5(2): 163-182.

Engel, R.S. and Henderson, S. (2013) 'Beyond Rhetoric: Establishing Police Academic Partnerships that Work', In Brown, J. (Ed.), *The Future of Policing*, London: Routledge, pp 217-236.

Engel, R.S. and Whalen, J.L. (2010) 'Police Academic Partnerships: Ending the Dialogue of the Deaf, the Cincinnati Experience', *Police Practice and Research, an International Journal,* 11(2): 105-116.

Evans, R. and Wilkinson, C. (1990) 'Variations in Police Cautioning Policy and Practice in England and Wales', *Howard Journal of Criminal Justice,* 29: 155-176.

Famega, C., Hinkle, J. and Weisburd, D.W. (2016) 'Why getting inside the 'Black Box' is important: Examining Treatment Implementation and Outputs in Police Experiments', *Police Quarterly,* August 25, DOI 20161098611116664336: 1-27.

Farrington, David P. (1983) 'Randomized experiments on crime and justice', In Tonry, M. and Morris, N. (Eds.) *Crime and Justice: An Annual Review of Research*. Volume IV. Chicago: University of Chicago Press, pp. 257-308.

Farrington, D.P. (2003) 'A Short History of Randomized Experiments in Criminology', *Evaluation Review,* 27(3): 218-227.

Farrington, D.P. and Joliffe, D. (2002) *A Feasibility Study into using a Randomised Controlled Trial to Evaluate Treatment Pilots at HMP Whitemoor*,

London: Home Office Research and Statistics Directorate, Online Report 14/02.

Farrington, D.P. and Welsh, B. (2006) 'A Half Century of Randomized Experiments on Crime and Justice', *Crime and Justice,* 34(1): 55-132.

Feder, L., Jolin, A. and Feyerherm, W. (2000) 'Lessons from two randomized experiments in criminal justice settings', *Crime and Delinquency,* 46(3): 380-400.

Federal Judicial Center Advisory Committee on Experimentation in the Law (1981) *Experimentation in the Law: Report of the Federal Judicial Center Advisory Committee.* Washington DC: Government Printing Office.

Feinman-Nemser, S. and Remillard, J. (1991) Perspectives on Learning to teach. In Murray, F.B. (Ed.) *The Teacher Educator's Handbook*. San Francisco: Jossey Bass: 63-91.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.F. and Wallace, F. (2005) *Implementation Research: A Synthesis of the Literature*, Tampa, FL: University of South Florida.

Flay, B.R., Biglan, A., Boruch, R.F., Gonzalez Castro, P., Gottfredson, D.C., Kellam, S., Moscicki, E.K., Schinke, S., Valentine, J.C. and Ji, P. (2005) 'Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination', *Prevention Science*, 6(3): 151-175.

Galliher, J.F. (1999) 'Against Administrative Criminology', *Social Justice,* 26(2): 56-59.

Gill, C.E., Wilson, D.B., Tuffin, R. and Weisburd, D.W. (2014) 'Foreword to the special issue: NPIA systematic reviews in policing', *Journal of Experimental Criminology,* 10(4): 367-369.

Giller, H. and Tutt, N. (1987) 'Police cautioning of juveniles: the continuing practice of diversity', *Criminal Law Review,* June: 367-374.

Gladwell, M. (2002) *The Tipping Point: How little things can make a big difference,* New York: Abacus.

Glaser, B.G. (1992) *Basics of Grounded Theory Analysis: Emergence vs. forcing*, Mill Valley, CA.: The Sociology Press.

Glaser, B.G. and Strauss, A. (1967) *The Discovery of Grounded Theory*, Chicago: Aldine.

Glennerster, R. and Takavarasha, K. (2013) *Running Randomized Evaluations: a practical guide*, Princeton: Princeton University Press.

George Mason Evidence-based Policing Matrix (2016) Retrieved 17th December 2016 from http://cebcp.org/evidence-based-policing/the-matrix/inclusion-criteria-methods-key/.

Goldacre, B. (2013) *Bad Pharma*, London: Harper Collins.

Goldacre, B., Drysdale H and Powell-Smith A, (2016) The COMPare Trials Project, Retrieved 17th December 2016 from www.COMPare-trials.org.

Goldman, J. (1983) Experimenting with Justice: the Federal Judicial Center Report, American Bar Foundation Research Journal, 733-739.

Gottfredson, D.C., Cook, T.D., Gardner, F.E.M., Gorman-Smith, D., Howe, G.W., Sandler, I.N. and Zafft, K.M. (2015) 'Standards for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation', *Prevention Science,* 16: 893-926.

Greene, J. (2014) 'New Directions in Policing: Balancing Prediction and Meaning in Police Research', *Justice Quarterly,* 31(3): 193-228.

Greenhalgh, T., Robert, G., MacFarlane, F., Bate, P. and Kyriakidou, O. (2004) 'Diffusion of Innovations in Service Organisations: Systematic Review and Recommendations', *Millbank Quarterly,* 82(4): 581-629.

Greenwood, P. and Turner, S. (1983) 'Evaluation of the Paint Creek Youth Center: a residential programme for serious delinquents', *Criminology,* 31(2), 263-279.

Grimshaw, J.M., Eccles, M.P., Lavis, J.N., Hill, S.J. and Squires, J.E. (2012) 'Knowledge Translation of Research Findings', *Implementation Science,* 7(50): 1-17. Retrieved 9th December 2016 from http://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-7-50.

Grol, R. and Grimshaw, J. (2003) 'From Best Evidence to Best Practice: Effective Implementation of Change in Patients' Care.' *Lancet,* 362(9391): 1225–1230.

Guest, G., Bunce, A. and Johnson, L. (2006) 'How many interviews are enough? An experiment with data saturation and variability', *Field Methods,* 18(1): 59-82.

Haas, N.E. (2016) 'Explaining Police Attitudes toward the Use of Force: The Case of Buenos Aires', In Mathieu Deflem (ed.) *The Politics of Policing: Between Force and Legitimacy (Sociology of Crime, Law and Deviance, Volume 21,* Bingley: Emerald Group Publishing Limited, pp.133 – 152.

Haines, A., Goldson, B., Haycox, A., Houten, R., Lane, S., McGuire, T., Perkins, N.E., Richards, S. and Whittington, R. (2012*) Evaluation of the Youth*

Justice Liaison and Diversion (YJLD) Pilot Program, Liverpool: University of Liverpool.

Halpern, D. (2017) *The Rise of Experimental Government,* Presentation to the What Works in Crime Reduction Conference at the British Library, 24th January 2017.

Hampshire Constabulary (1982) 'Evaluation of the relative effectiveness of instant and deferred juvenile cautions' [Unpublished paper for Hampshire Constabulary].

Hampshire Constabulary (1988) 'A Scrutiny into Community Relations'. [Unpublished report for Hampshire Constabulary].

Hawken, A. and Kleiman, M. (2009) *Managing drug-involved probationers with swift and certain sanctions: evaluating Hawaii's HOPE. Evaluation Report*, NCJ 229023, Washington DC: National Institute of Justice.

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012) *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*, London: Cabinet Office, Behavioural Insights Team.

Heitjan, D.F. (1999) 'Causal Inference in a Clinical Trial: a comparative example', *Controlled Clinical Trials,* 20: 309-318.

Hesse-Biber, S.N., Dupuis, B. and Kinder, S. (1991) 'HyperRESEARCH: A computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis', Qualitative Sociology, 14(4): 289-306.

Higgins, J.P.T. and Altman, D.G. (2008) 'Assessing risk of bias in included studies', In Higgins, J.P.T. and Green, S. (Eds.) (2008) *Cochrane Handbook*

for Systematic Review Interventions, Chichester: Wiley, Cochrane Book Series, pp 187-242.

Higgins, J.P.T. and Green, S. (Eds.) (2008) *Cochrane Handbook for Systematic Review Interventions*, Chichester: Wiley, Cochrane Book Series.

Higginson, A., Eggins, E., Mazerolle, L. and Stanko, E. (2014) *The Global Policing Database [Database and Protocol]*, Retrieved 5th December 2016 from http://www.gpd.uq.edu.au/search.php.

HIrschi, T. (1993) 'Administrative Criminology', *Contemporary Sociology,* 22: 348-50.

Hobday, J. (2015). 'Targeting Reasons for Rejecting Random Assignment in an RCT', [Unpublished Dissertation submitted for Master's degree in Applied Criminology and Management, Institute of Criminology, University of Cambridge].

Hollis, S. and Campbell, F. (1999) 'What is meant by intention to treat analysis? Survey of published randomised controlled trials', *British Medical Journal*, 319: 670-4.

Holman, C., Piper, S.K., Grittner, U., Diamantaras, A.A., Kimmelman, J., Siegerink, B. and Dirnagl, U.(2016) 'Where Have All the Rodents Gone? The Effects of Attrition in Experimental Research on Cancer and Stroke', *PLoS Biology*, 14(1): e1002331.doi:10.1371/journal.pbio.1002331.

Home Office (2015) *Integrated Offender Management Guidance*, London: Home Office. Retrieved 9th December 2016 from https://www.gov.uk/guidance/integrated-offender-management-iom.

Hope, T. (2009) 'The Illusion of Control: A Response to Professor Sherman', *Criminology and Criminal Justice,* 9(2): 125-134.

Hough, M. (2010) ‘Gold standard or fool’s good? The pursuit of certainty in experimental criminology’, *Criminology and Criminal Justice,* 10(1): 11-23.

Huey, L. and Mitchell, R.J. (2016) ‘Unearthing Hidden Keys: Why Pracademics Are an Invaluable (If Underutilized) Resource in Policing Research’, *Policing,* 10(3): 300-307.

ICPR (Institute for Criminal Policy Research) (2012) *Assessing Young People in Police Custody: An Examination of the operation of Triage Programs,* London: Home Office, Occasional Paper 106.

Irving, B. and Dixon, B. (2002) '*Hotspotting': Turning Police Theory into Practice in Thames Valley and Northumbria,* London: The Police Foundation.

Iszatt-White, M. and Saunders, C. (2014) *Leadership*, Oxford: Oxford University Press.

Juni, P., Altman, D.G. and Egger, M. (2001) ‘Systematic Reviews in health care: Assessing the quality of controlled clinical trials’, *British Medical Journal,* 323: 42-46.

Juni, P., Witschi, A., Bloch, R. and Egger, M. (1999) ‘The hazards of scoring the quality of clinical trials for meta-analysis’, *Journal of the American Medical Association,* 282: 1054-1060.

Kahneman, D. (2011) *Thinking, fast and slow*, London: Allen Lane.

Kawamoto, K., Houliahan, C., Balas, E. and Lobach, D. (2005) ‘Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success.’ *British Medical Journal,* 330(7494): 765.

Kelling, G., Pate, T., Dieckman, D. and Brown, C.E. (1974) *The Kansas City Preventive Patrol experiment: technical report*, Washington, DC: Police Foundation.

Kerr, J., Ranns, H., Ludford, H., Barnard, M., Purdon, S., Barrett, B. and Farrington, D.P. (2011) *Assessing the feasibility of conducting a randomised controlled trial or other outcome study of P-ASRO*, London: Ministry of Justice, National Offender Management Service.

Konze, A. (2016) 'An update on the survey of European Law Enforcement Education Systems (ELEES)', [Paper presented to the CEPOL European Police Research and Science Conference, October 7th 2016].

Koper, C. (1995) 'Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hotspots', *Justice Quarterly,* 12(4): 649-672.

Kotter, J.P. (2012) *Leading Change*, 1st. Revised Edition, Cambridge, MA: Harvard Business Review Press.

Landau, S.F. and Nathan, G. (1983) 'Selecting Delinquents for cautioning in the London Metropolitan Area', *British Journal of Criminology,* 23(2): 128-149.

Laycock, G. and Mallender, J. (2015) 'Right method, right price: the economic value and associated risks of experimentation', *Journal of Experimental Criminology*, 11: 653-658.

Laycock, G. and Tarling, R. (1985) 'Police Force Cautioning: Policy and Practice', *The Howard Journal of Criminal Justice*, 24: 81-92.

Lee, Y. J., Ellenberg, J. H., Hirtz, D. G. and Nelson, K. B. (1991) 'Analysis of clinical trials by treatment actually received: Is it really an option?', *Statistics in Medicine,* 10: 1595–1605.

Lewins, A. and Silver, C. (2007) *Using software in Qualitative Research: A step-by-step Guide*, London: Sage.

Lipsey, M.W. and Wilson, D.B. (2001) *Practical Meta-Analysis*, Thousand Oaks, CA: Sage.

Loftus, B. (2009) *Police Culture in a Changing World*, Oxford: Oxford University Press.

Lum, C. and Koper, C. (2017) *Evidence-based policing: translating research into practice.* Oxford: Oxford University Press.

Lum, C., Koper, C. and Telep, C.W. (2011) 'The Evidence-based Policing Matrix', *Journal of Experimental Criminology*, 7: 3-26.

Lum, C., Telep, C., Koper, C. and Grieco, J. (2012) 'Receptivity to Research in Policing', *Justice, Research and Policy,* 14(1): 61-95.

Lum, C., Koper, C.S., Merola, L.M., Scherer, A., and Reioux, A. (2015) *Existing and Ongoing Body Worn Camera Research: Knowledge gaps and opportunities*, Report for the Laura and John Arnold Foundation, Fairfax, VA: Center for Evidence-Based Crime Policy, George Mason University.

McGloin, J.M. and Thomas, K.J. (2013) 'Experimental tests of Criminological Theory', In Welsh, B.C., Braga, A.A. and Bruinsma, G.J.N. (Eds.) *Experimental Criminology: Prospects for advancing science and public policy*, Cambridge: Cambridge University Press, pp. 15-43.

Mackenzie, D.L. (2006) *What works in corrections: reducing the criminal activities of offenders and delinquents,* New York: Cambridge University Press.

MacQueen, S. and Bradford, B. (2016) 'Where did it all go wrong? Implementation failure - and more - in a field experiment of procedural justice

policing', *Journal of Experimental Criminology, Online First: DOI 10.1007/s11292-016-9278-7*.

Maguire, E.R., Uchida, C.D. and Hassell, K.D. (2015) 'Problem-Oriented Policing in Colorado Springs: A Content Analysis of 753 Cases', *Crime and Delinquency,* 61(1): 71-95.

Manning, P. (2005) 'The study of policing', *Police Quarterly,* 8(1): 23–4.

Manning, P. (2011) *Democratic Policing in a Changing World*, Boulder, CO: Paradigm Press.

Marks, M. and Sklansky, D. (Eds.) (2012) *Police Reform from the bottom up: Officers and their Unions as agents of change*, Abingdon: Routledge.

Martinson, R. (1974) 'What works? Questions and answers about prison reform', *The Public Interest,* Spring: 22-54.

Maxfield, M.G. and Babbie, E.R. (2015) *Research Methods for Criminal Justice and Criminology*, Seventh Edition. Stamford, CT: Cengage Learning.

Mazerolle, L.W., Lum, C. and Braga, A.A. (2014) 'Using Experimental Designs to Study Police Interventions', In Reisig, M. and Kane, R.J. (Eds.) *Police and Policing*, Oxford: OUP, pp. 487-518.

Miles, M.B. and Huberman, A.M. (1994) *Qualitative Data Analysis: An expanded sourcebook*, Newbury Park, CA.: Sage.

Ministry of Justice and National College of Policing (2013) *Consultation on Out of Court Disposals*, London: Ministry of Justice and National College of Policing, Retrieved 9th December 2016 from https://consult.justice.gov.uk/digital-communications/out-of-court-disposals/supporting_documents/outofcourtdisposalsconsultation.pdf.

Ministry of Justice and Youth Justice Board (2013) *Modern Youth Offending Partnerships: Guidance on Effective Youth Offending Team Governance in England*, London: Ministry of Justice and Youth Justice Board, Retrieved 9th December 2016 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/319291/youth-offending-partnerships-guidance.pdf.

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P. and Walsh, S. (2005) 'Assessing the quality of randomised controlled trials: An annotated bibliography of scales and checklists', *Controlled Clinical Trials,* 16: 62-73.

Moreno, Y. and Hughes, P. (2008) *Effective Prosecution: Working in partnership with the CPS,* Oxford: Oxford University Press.

Moore, S., Sellen, J., Crompton, K. and Hallingberg, B. (2012) *A Brief evaluation of Cardiff Triage*, Cardiff: Cardiff University.

Morse, J.M. (2007) 'Sampling in Grounded Theory', In Bryant, A. and Charmaz, K. (Eds.) *The Sage Handbook of Grounded Theory*, Thousand Oaks, CA.: Sage, pp. 229-245.

Mott, J. (1983) 'Police Decisions for dealing with juvenile offenders', *British Journal of Criminology,* 23(8): 249-270.

National Research Council (2009) *Strengthening Forensic Science in the United States: A Path Forward,* Washington, DC: The National Academies Press.

Neyroud, E.C. (2015) 'The prevalence, incidence, type and harm of police reported victimization of offenders in the Turning Point Project', [Unpublished Dissertation submitted for degree of Master of Philosophy, Institute of Criminology, University of Cambridge].

Neyroud, P.W. (2011a) *Operation Turning Point: an experiment in 'offender-desistance policing': Crimport Protocol*, Retrieved 2nd December 2016 from http://www.crim.cam.ac.uk/research/experiments/rex-post/operation_turning_point.pdf.

Neyroud, P.W. (2011b) *Review of Police Leadership and Training*, London: Home Office.

Neyroud, P.W. (2015a) 'Future Perspectives in Policing: A Crisis or a Perfect Storm: The Trouble with Public Policing', In Wankhade, P. and Weir, D. (Eds.) *Police Services: Leadership and Management Perspectives,* Dordrecht: Springer, pp. 161-167.

Neyroud, P.W. (2015b) *Rethinking the Gateway: Using Evidence to reform the Criminal Justice System for victims and people who offend*, London: Howard League 'What if?" series. Retrieved 9th December 2016 from http://howardleague.org/wp-content/uploads/2016/03/Rethinking-the-gateway.pdf.

Neyroud, P.W. (2015) 'Evidence-based triage in prosecuting arrestees: Testing an actuarial system of selective targeting', *International Criminal Justice Review*, 25: 117-131.

Neyroud, P.W. (2017) 'Researching the Police: Inside-Outside Perspectives in a new world of police professionalism and practitioner research', In Cowburn, M., Gelsthorpe, L. and Wahidin, A. (Eds.) *Research Ethics in Criminology: Dilemmas, Issues and Solutions*, London: Routledge, Chapter 5, 77-94.

Neyroud, P.W. and Sherman, L.W. (2013) 'Dialogue and Dialectic: Police Legitimacy and the new Professionalism', In Tankebe, J. and Liebling, A. (Eds.) *Legitimacy and Criminal Justice: An International Exploration*, Oxford: OUP, pp. 293-308.

Neyroud, P.W. and Slothower, M. (2012) 'Operation Turning Point: the first interim report', [Unpublished report, Cambridge: Institute of Criminology].

Neyroud, P.W. and Slothower, M. (2013) 'Operation Turning Point: the second interim report', [Unpublished report, Cambridge: Institute of Criminology].

Neyroud, P.W. and Slothower, M.P. (2015), 'Wielding the Sword of Damocles: The Challenges and Opportunities in Reforming Police Out-of-Court Disposals in England and Wales', In Wasik, M. and Santatzoglou, S. (Eds.) *The Management of Change of Criminal Justice: Who knows best?,* Basingstoke: Palgrave MacMillan, pp. 275-293.

Neyroud, P.W. and Weisburd, D. (2014a) 'Transforming the Police through Science: some new thoughts on the controversy and challenge of translation', *Translational Criminology,* Spring: 16-18.

Neyroud, P.W. and Weisburd, D.W. (2014b) 'Transforming the Police Through Science: The Challenge of Ownership', *Policing*, 8(4): 287-293.

NHS (2015) *Clinical Trials*, Retrieved 2nd December 2016 from http://www.nhs.uk/Conditions/Clinical-trials.

Nutley, S.M., Walter, I. and Davies, H.T.O. (2007) *Using Evidence: How research can inform public service*, Bristol: Policy Press.

Open Science Collaboration (2015) 'Estimating the Reproducibility of Psychological Science' *Science,* 349(6251): DOI: 10.1126/science.aac4716.

Patsapoulos, N. (2011) 'A Pragmatic view of Pragmatic Trials', *Dialogues in Clinical Neuroscience,* 13(2): 217-224.

Patterson, G.T., Chung, I.W. and Swan, P.G. (2012) *The effects of stress management interventions among police officers and recruits,* Campbell Systematic Review Library. Retrieved 6th December 2016 from

https://www.campbellcollaboration.org/library/download/27_1ca66e1e1f4d6c2f095a91398f1c5d42.html.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage.

Perry, A.E., Weisburd, D.W. and Hewitt, C. (2010) 'Are Criminologists describing randomised controlled trials in ways that allow us to assess them? Findings from a sample of crime and justice trials', *Journal of Experimental Criminology*, 6: 245-262.

Petrosino, A.J. (1997) 'What works revisited again: A meta-analysis of Randomized Field Experiments in Individual-level interventions', [Unpublished PhD Thesis, Rutgers University].

Petrosino, A.J., Boruch, R.F., Rounding, C., McDonald, S. and Chalmers, I. (2000) 'The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to Facilitate the Preparation and Maintenance of Systematic Reviews of Social and Educational Interventions.', *Evaluation & Research in Education*, 14(3-4): 206-219.

Petrosino, A.J., Kiff, P. and Lavenberg, J.G. (2006) 'Research Note: Randomized Field Experiments Published in the British Journal of Criminology, 1960-2004', *Journal of Experimental Criminology,* 2(1): 99-111.

Petrosino, A.J. and Soydan, H. (2005) 'The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research', *Journal of Experimental Criminology,* 1(4): 435-450.

Petrosino, A., Turpin-Petrosino, C. and Guckenburg, S. (2010) *Formal System Processing of Juveniles: effects on delinquency,* Campbell Collaboration (Crime and Justice) Systematic Review. Retrieved 9th December 2016 from https://www.campbellcollaboration.org/library/formal-system-processing-of-juveniles-effects-on-delinquency-a-systematic-review.html.

Porter, G. (2015) 'A protocol and experimental trial: The Checkpoint Desistance Programme in Durham' [Unpublished M.St. thesis, Institute of Criminology, University of Cambridge.

President's Task Force on Victims of Crime (1982) *Final Report*, Retrieved 9th December 2016 from https://www.ovc.gov/publications/presdntstskforcrprt/87299.pdf.

President's Task Force on 21st Century Policing (2015) *Final Report of the President's Task Force on 21st Century Policing*, Washington, DC: Office of Community Oriented Policing Services.

Punch, K.F. (2005) *Introduction to Social Research: Quantitative and Qualitative Approaches*, London: Sage.

Quay, H.C. (1977) 'The Three Faces of Evaluation: What can be expected to work', *Criminal Justice and Behavior,* 4(4): 341-354.

Reiner, R. (2010) *The Politics of the Police*, 4th Edition, Oxford: Oxford University Press.

Reuland, M. (2004) *A Guide to Implementing Police-based Diversion Programs for People with Mental Illness,* Delmar, NY: Technical Assistance and Policy Analysis Center for Jail Diversion.

Rice, L. (2010) *Conditional Cautions: lessons learnt from the unpaid reparative work pilot implementation,* London: Ministry of Justice Research Summary 5/10.

Riecken, H.W., Boruch, R.F., Campbell, D.T., Caplan, N., Glennanu, T.K., Pratt, J.W., Rees, A. and Williams, W.W. (1974) *Social Experimentation: a method for planning and evaluating social programs*, New York: Academic Press.

Rogers, E. M. (1995) *Diffusion of innovations*, 4th edition., New York: Free Press.

Rogers, E.M. (2002) 'Diffusion of Preventive Innovations', *Addictive Behaviors,* 27: 989-993.

Rojek, J., Martin, P. and Alpert, G.P. (2015) *Developing and Maintaining Police-Researcher Partnerships to Facilitate Research Use: A comparative analysis*, New York: Springer.

Rousseau, D. M. (2012) 'Envisioning Evidence-Based Management', In Rousseau, D.M. (Ed.) *The Oxford Handbook of Evidence-Based Management.* New York: Oxford University Press, pp. 3-25.

Rynes, S. (2007) 'Let's create a Tipping Point: What academics and practitioners can do, alone and together', *Academy of Management Journal,* 50(5): 1046-1054.

Sampson, R.J. (2010). Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology. *Journal of Quantitative Criminology, 25*, 489-500.

Sampson, R.J. and Laub, J.H. (1993) *Crime in the Making: Pathways and Turning Points through Life,* Cambridge Mass.: Harvard University Press.

Sampson, R.J. and Laub, J.H. (2003) *Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70,* Cambridge Mass.: Harvard University Press.

Sampson, R.J., Winship, C. and Knight, C. (2013) 'Translating causal claims principles and strategies for policy-relevant criminology', *Criminology and Public Policy,* 12(4): 587-616.

Sandars, A. (1988) ‘The Limits to diversion from prosecution’, *British Journal of Criminology,* 28(4): 513-532.

Santos, R.B. and Santos, R.G. (2015) ‘Examination of police dosage in residential burglary and residential theft from vehicle micro-time hot spots’, *Crime Science,* 4(27): 1-12.

Schulz, K. F., Chalmers, I., Hayes, R. J. and Altman, D. G. (1995) ‘Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials’, *Journal of American Medical Association,* 273: 408–412.

Schwalbe, C.S., Gearing, R.E., MacKenzie, M.J., Brewer, K.B., Ibrahim, R. (2012) ‘A meta-analysis of experimental studies of diversion programs for juvenile offenders’, *Clinical Psychology Review,* 32(1): 26-33.

Schwartz D. and Lellouch J. (1967) ‘Explanatory and pragmatic attitudes in therapeutical trials’, *Journal of Chronic Diseases,* 20: 637–648.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Belmont, CA.: Wadsworth Cengage Learning.

Shapland, J., Robinson, G. and Sorsby, A. (2011) *Restorative Justice in Practice: Evaluating What Works for Victims and Offenders*, Abingdon: Routledge.

Shepherd, J.P. (2002) ‘Public services need practitioner-academics’, *Times Higher Educational Supplement,* 9 August: 14.

Shepherd, J.P. (2003) ‘Explaining Feast or Famine in Randomized Field Trials: Medical Science and Criminology compared’, *Evaluation Review,* 27(3): 290-315.

Shepherd, J.P. (2015) *Professionalising the Probation Service: Why University Institutes would transform rehabilitation*, London: Howard League "What if?" series. Retrieved 9th December 2016 from http://howardleague.org/wp-content/uploads/2016/04/Professionalising-the-probation-service.pdf.

Sherman, L.W. (1979) 'The case for the research police department' . *Police Magazine,* 2 (6): 58-59

Sherman, L.W. (1984) 'Experiments in Police Discretion: Scientific Boon or Dangerous Knowledge?', *Law and Contemporary Problems,* 47(4): 61-81.

Sherman, L.W. (1992) *Policing Domestic Violence: Experiments and Dilemmas*, New York: Free Press.

Sherman, L.W. (1997) 'Policing for Crime Prevention', In Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. and Bushway, S. (1997) *Preventing Crime: What Works, What Doesn't, What's Promising,* Washington, DC: Office of Justice Programs, Chapter 8: pp. 1-58.

Sherman, L.W. (1998) *Evidence-based policing. Ideas in American Policin*g, Washington, DC: Police Foundation.

Sherman, L.W. (2009) 'Evidence and Liberty: the promise of experimental criminology', *Criminology and Criminal Justice,* 9: 5-28.

Sherman, L.W. (2010) 'An Introduction to Experimental Criminology', In Piquero, A.R. and Weisburd, D. (Eds.) *Handbook of Quantitative Criminology*, New York: Springer, pp. 399-437.

Sherman, L. W. (2011a) 'Al Capone, the Sword of Damocles, and the Police–Corrections Budget Ratio', *Criminology & Public Policy*, 10: 195-206.

Sherman, L.W. (2011b) ‘Offender Desistance Policing (ODP): less prison and more evidence in rehabilitating offenders’, In Bliesener, T., Beelman, A. and Stemmler, M. (Eds.) *Anti-Social Behavior and Crime: contributions of developmental and evaluation research to prevention and intervention*, Cambridge, MA: Hogrefe Publishing, pp. 199-218.

Sherman, L.W. (2013) ‘The Rise of Evidenced-Based Policing: Targeting, Testing and Tracking’ In Tonry, M. (Ed.) *Crime and Justice in America 1975-2025*, Chicago: University of Chicago Press, Crime and Justice, 42.

Sherman, L.W., Farrington, D.P., Welsh, B. and MacKenzie, D. (Eds.) (2002) *Evidence-based Crime Prevention*, London: Routledge.

Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. and Bushway, S. (1997) *Preventing Crime: What Works, What Doesn’t, What’s Promising,* Washington, DC: Office of Justice Programs.

Sherman, L.W. and Neyroud, P.W. (2011) ‘Offender Desistance Policing (ODP): an outline proposal for research’, [Unpublished Paper for West Midlands Police Chief Constable. Cambridge: Institute of Criminology, University of Cambridge (DOC/TP/1)].

Sherman, L.W. and Neyroud, P.W. (2012) *Offender Desistance Policing and the Sword of Damocles*, London: Civitas.

Sherman, L.W. and Strang, H.S. (2004) ‘Verdicts or Inventions? Interpreting Results from Randomized Controlled Experiments in Criminology’, *American Behavioral Scientist,* 47(5): 575-607.

Sherman, L.W., Williams, S., Ariel, B., Strang, L.R., Wain, N., Slothower, M.P. and Norton, A. (2014) ‘An Integrated Theory of Hot Spots Patrol Strategy: Implementing Prevention by Scaling Up and Feeding Back’, *Journal of Contemporary Criminal Justice,* 30: 95-122.

Skinns, L., Wooff, A. and Sprawson, A. (2016) 'The ethics of researching the police: dilemmas and new directions' In Brunger, M., Tong, S. and Martin, D. (Eds.) *Introduction to Policing Research: taking lessons from practice*, Abingdon: Routledge, pp. 185-203.

Skogan, W. (2008) 'Why Reforms Fail', *Policing and Society,* 18(1): 23-34.

Skogan, W. and Frydl, K. (2004) *Fairness and Effectiveness in Policing*, Committee to Review Research on Police Policy and Practices. Washington, DC: The National Academies Press.

Slothower, M.P. (2014b) 'Strengthening Police Professionalism with Decision Support: Bounded Discretion in Out-of-Court Disposals', *Policing,* 8(4): 353-367.

Slothower, M.P., Sherman, L.W. and Neyroud, P.W. (2015) 'Tracking Quality of Police Actions in a Victim Contact Program: A Case Study in Training, Tracking and Feedback (TTF) in Evidence-based Policing', *International Criminal Justice Review,* 25(1): 98-116.

Smith, D. (2012) 'Arguments about methods in Criminal Justice Evaluations', In Bowen, E. and Brown, S. (Eds.) *Perspectives on Evaluating Criminal Justice and Corrections,* Bingley: Emerald Group Publishing Limited, pp. 49-71.

Sorg, E.T, Wood, J.T., Groff, E.R. and Ratcliffe, J.H. (2016) 'Explaining Dosage Diffusion During Hot Spot Patrols: An Application of Optimal Foraging Theory to Police Officer Behavior', *Justice Quarterly, Online First*. Retrieved 11th December 2016 from: http://www.tandfonline.com/action/showCitFormats?doi=10.1080/07418825.2016.1244286.

Sosa, K. (2012) *Proceed with Caution: use of out of court disposals in England and Wales,* London: Police Exchange.

Sparrow, M.K. (2011) *Governing Science: New Perspectives in Policing*, Washington, DC: Department of Justice, National Institute of Justice.

Sparrow, M.K. (2016) *Handcuffed: What holds Policing Back, and the Keys to Reform,* Washington, DC: Brooking Institution Press.

Spoth, R., Rohrbach, L.A., Greenberg, G., Leaf, P., Hendricks Brown, C., Fagan, A, Catalano, R.F., Pentz, M.-A., Sloboda, Z., Hawkins, J.D. and Society for Prevention Research Type 2 Translational Task Force Members and Contributing Authors (2013) 'Addressing Core Challenges for the Next Generation of Type 2 Translation Research and Systems: The Translation Science to Population Impact (TSci Impact) Framework', *Prevention Science,* 14: 319-351.

Stith, S., Pruitt, I., Dees, J., Fronce, M., Green, N., Som, A. and Linkh, D. (2006) 'Implementing community-based prevention programming: A review of the literature', *Journal of Primary Prevention,* 27: 599–617.

Stone, C. and Travis, J. (2011) *Towards a new Professionalism in Policing. New Perspectives in Policing,* Washington, DC: Department of Justice, National Institute of Justice.

Strang, H. (2012) 'Coalitions for a common purpose: managing relationships in experiments', *Journal of Experimental Criminology,* 8(3): 211-227.

Strang, H., Sherman, L.W., Mayo-Wilson, E., Woods, D. and Ariel, B. (2013) *Restorative Justice Conferencing: Effects of Face-to-Face Meetings of Offenders and Victims,* Campbell Collaboration (Crime and Justice) Systematic Review. Retrieved 9th December 2016 from https://www.campbellcollaboration.org/library/restorative-justice-conferencing-rjc-using-face-to-face-meetings-of-offenders-and-victims-effects-on-offender-recidivism-and-victim-satisfaction-a-systematic-review.html.

Strauss, A. (1987) *Qualitative Analysis for Social Scientists*, New York: Cambridge University Press.

Strauss, A. and Corbin, J. (1990) *Basics of Qualitative Research*, Newbury Park, CA.: Sage.

Test, M.F. and Burke, S. (1985) 'Random assignment of chronically ill persons to hospital or community treatment', *New Directions for Program Evaluation,* 28: 81-94.

Tilley, N. (2009) 'Sherman vs Sherman: Realism vs Rhetoric', *Criminology and Criminal Justice,* 9(2): 135-144.

Torgerson, C.J. (2003) *Systematic Reviews*, London: Continuum.

Torgerson, D.J. and Torgerson, C.J. (2008) *Designing Randomised Controlled Trials in Health, Education and Social Sciences: An Introduction*, Basingstoke: Palgrave MacMillan.

Tyler, T. (Ed.) (2007) *Legitimacy and Criminal Justice,* New York: Russell Sage Foundation.

Urquhart, C. (2013) *Grounded Theory for Qualitative Research: A Practical Guide*, London: Sage.

United Nations (1985) *Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power*, Retrieved 9th December 2016 from http://www.un.org/documents/ga/res/40/a40r034.html.

Van Dijk, A., Hoogewonning, F. and Punch, M. (2015) *What matters in policing? Change, Values and Leadership in Turbulent Times*, Bristol: Policy Press.

Vito, G., Walsh, W.F. and Kunselman, J. (2005) 'Community Policing: The Middle Manager's Perspective' *Police Quarterly,* 8: 490-511.

Wain, N. and Ariel, B. (2014) 'Tracking of Police Patrol', *Policing,* 8(3): 274-283.

Walker, S. (1993) 'Does Anyone Remember Team Policing? Lessons for the Team Policing Experience for Community Policing', *American Journal of Police,* 12(1): 33-55.

Walumba, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S. and Peterson, S. J. (2008) 'Authentic leadership: development and validation of a theory-based measure', *Journal of Management,* 34(1): 89–126.

Washington State Institute for Public Policy (2016) *Benefit-Cost Technical Documentation,* Retrieved 22nd January 2017 from http://www.wsipp.wa.gov/TechnicalDocumentation/WsippBenefitCostTechnicalDocumentation.pdf.

Weatheritt, M. (1986) *Innovations in Policing*, London: Police Foundation.

Weisburd, D.W. (2003) 'Ethical practice and evaluation of interventions in crime and justice: the moral imperative for randomised trials', *Evaluation Review*, 27: 336-354.

Weisburd, D.W. (2005). 'Hot Spots Policing Experiments and Criminal Justice Research: Lessons from the Field', *The Annals of the American Academy,* 599: 220-245.

Weisburd, D.W. and Hinkle, J.C. (2012) 'The Importance of Randomised Experiments in Evaluating Crime Prevention', In Welsh, B.C. and Farrington, D.P. (Eds.) *The Oxford Handbook of Crime Prevention*, Oxford: OUP, pp. 446-466.

Weisburd, D.W. and Neyroud, P.W. (2011) *Police Science: Towards a new Paradigm*, Washington DC: Department of Justice, National Institute of Justice.

Weisburd, D.W., Petrosino, A.J. and Mason, G. (1993) *When bigger is not better: Design sensitivity in a sample of criminal justice experiments*, Washington, DC: National Institute of Justice.

Weisburd, D.W., Sherman, L.W. and Petrosino, A.J. (1990) *Registry of Randomized Criminal Justice Experiments in Sanctions*, Washington, DC: National Institute of Justice.

Weisburd, D.W. and Telep, C. (2015) *The Efficiency of Place-based Policing*, Hebrew University of Jerusalem, Legal Studies Research Paper No. 15-26.

Welsh, B.C. and Farrington, D.P. (2012) 'Crime Prevention and Public Policy', In Welsh, B.C. and Farrington, D.P. (Eds.) *The Oxford Handbook of Crime Prevention*. Oxford: OUP, pp. 3-23.

Welsh, B.C., Braga, A.A., & Bruinsma, G.J.N. (2013a) 'New perspectives and developments in experimental criminology', *Policing: A Journal of Policy and Practice*, 7: 411-418.

Welsh, B.C., Braga, A.A., & Bruinsma, G.J.N. (2013b) 'Experimenting with Crime and Criminal Justice', In Welsh, B.C., Braga, A.A., & Bruinsma, G.J.N (Eds.). *Experimental Criminology*. Cambridge: Cambridge University Press, pp. 1-25.

Willis, J.J. (2016) 'The Romance of Police Pracademics', Policing, 10(3): 315-321.

Wilson, O.W. (1950) *Police Administration*, New York: McGraw-Hill.

Wolfe, S. and Piquero, A. (2011) 'Organizational Justice and Police Misconduct', *Criminal Justice and Behavior,* 38(4): 322-353.

Wood, S., Eckley, L., Stuart, J., Hughes, K., Kelly, D., Harrison, D. and Quigg, Z. (2011) *Evaluation of the Youth Crime Action Plan (YCAP) in Liverpool,* Liverpool: Centre for Public Health, Liverpool John Moores University.

Wood, J. D., Sorg, E. T., Ratcliffe, J. H., Groff, E. R., & Taylor, C. J. (2014) 'Cops as treatment providers: Realities and ironies of police work in a foot patrol experiment', *Policing and Society*, 24: 362-379.

Yin, R.K. (2014) *Case Study Research: Design and Methods*, 5th Edition, Thousand Oaks: Sage.

7.2 Bibliography of Police Randomised Controlled Trials:

Note: Police RCTs listed here and marked with an asterisk (*) were published after the end of the time limit for the search in Chapter 1. Those indicated by two asterisks (**) are duplicate reports of the same RCT or group of RCTs.

Abrahamse, A. F., Ebener, P. A., Greenwood, P. W., Fitzgerald, N., & Kosin, T. E. (1991) 'An experimental evaluation of the Phoenix repeat offender program', *Justice Quarterly,* 8(2): 141-168.

Ackerley, D.G. (1986) 'The Effects of a stress management program on Police Personnel', [Unpublished PhD. thesis, University of Missouri-Columbia. Retrieved 7th December 2016 from http://fulfillment.umi.com/dissertations/003f6b839fdd6fcd19dc8a54b6df3b5f/1454168315/8716685.pdf.

Amendola, K.L, Weisburd, D., Hamilton, E.E., Jones, G, and Slipka, M. (2011) 'An experimental study of compressed work schedules in policing: advantages and disadvantages of various shift lengths', *Journal of Experimental Criminology,* 7: 407-442.

*Andersen, J. P., Gustafsberg, H. (2016) 'A training method to improve police use of force decision making: A randomized controlled trial', *SAGE Open*, 6. doi:10.1177/2158244016638708 (2 RCTs in Finland and Canada).

Angel, C., Sherman, L.W., Strang, H.S., Ariel, B., Bennett, S., Inkpen, N., Keane, A., Richmond, T. (2014) 'Shot-term effects of restorative justice conferences on post-traumatic stress symptoms among robbery and burglary victims: a randomized controlled trial', *Journal of Experimental Criminology,* 10(3): 291-307.

Antrobus, E. and Pilotto, A. (2016) 'The Queensland Unlawful Entry Experiment', *Journal of Experimental Criminology,* 12(3): 319–345.

**Ariel, B., Farrar, W.A. and Sutherland, A. (2014) 'The Effect of Police Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police: A Randomized Controlled Trial', *Journal of Quantitative Criminology,* 31(3): 509-535.

*Ariel, B., Smallwood, J., Sherman, L.W., Wain, N., Weinborn, C., Goodhill, W., Sosinski, G., Tankebe, J. and Yahalom, O. (2015). 'The Birmingham Hotspots Experiment: Operation Savvy', [Presentation to the 7th International Conference on Evidence-based Policing, University of Cambridge. Retrieved 8th December 2016 from http://www.crim.cam.ac.uk/events/conferences/ebp/2014/slides/223%20-%20WED%20-%20PCs%20vs%20PCSO%20-%20Barak%20Ariel.pdf.

*Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., Megicks, S. and Henderson, R. (2016b) 'Report: Increases in police use of force in the presence of body worn cameras are driven by officer discretion a protocol-based sub-group analysis of ten randomized experiments', *Journal of Experimental Criminology,* 12(3): 453-463

*Ariel, B., Weinborn, C. and Sherman, L.W. (2016a) '"Soft" policing at hot spots—do police community support officers work? A randomized controlled trial', *Journal of Experimental Criminology*, 12(3): 277-317.

Banerjee, A., Duflo, E., Keniston, D. and Singh, N. (2013) *Randomised Drunk Driving Crackdown in India*, Retrieved 7th December 2016 from http://keniston.commons.yale.edu/files/drunk_driving_03.2.pdf.

Banerjee, A., Raghabendra Chattopadhyay, Duflo, E., Keniston, D. and Singh, N. (2012) *Can Institutions be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police*, NBER Working Paper No. 17912. Retrieved 7th December 2016 from http://economics.mit.edu/files/7783.

Bennett, S. and Newman, M. (2015) ‘Mobile Community Police Office Experiment (Queensland)’, [Unpublished report for Queensland Police, Brisbane: University of Queensland and Queensland Police Service].

Berk, R., Campbell, A., Western, B. and Klap, R. (1992) ‘Colorado Springs Spouse Abuse Experiment’, *The Journal of Criminal Law and Criminology,* 83: 170-200.

Binder, A. and Newkirk, M. (1977) ‘University of California Youth Services Program’, *Crime Prevention Review,* 4: 26-32.

Boyanowsky, E.O. and Griffiths, C.T. (1982) ‘Weapons and Eye Contact as Instigators or Inhibitors of Aggressive Arousal in Police Citizen Interaction’, *Journal of Applied Social Psychology,* 12(4): 398-407.

**Bradford, B. and MacQueen, S. (2014) *The Scottish Community Engagement Trial (ScotCET): Interim report*, Edinburgh: Scottish Institute for Policing Research.

Braga, A. A., & Bond, B. J. (2008) ‘Policing crime and disorder hot spots: A randomized, controlled trial’, *Criminology,* 46(3): 577-607.

Braga, A. A., Weisburd, D. W., Waring, E. J., Mazerolle, L. G., Spelman, W., & Gajewski, F. (1999) ‘Problem-oriented policing in violent crime places: A randomized controlled experiment’, *Criminology,* 37(3): 541-580.

**Buerger, M.E. (1993) ‘Convincing the recalcitrant: re-examining the Minneapolis RECAP experiment’, [Unpublished PhD. Thesis, Rutgers University].

Byles, J.A. and Maurice, A. (1979) ‘The juvenile services project: an experiment in delinquency control’, *Canadian Journal of Criminology*, 21: 155-165.

Clayton, R. R., Cattarello, A. M., & Johnstone, B. M. (1996) 'The effectiveness of Drug Abuse Resistance Education (Project DARE): 5-year follow-up results', *Preventive Medicine,* 25(3): 307-318.

Davidson, W.S., Seidman, E., Rappaport, J., Berck, P.L., Rappaport, N. and Herring, J. (1977) 'The Urbana-Champaign Pre-court diversion trial', *Social Work Research and Abstracts,*13: 40-49. (Two RCTs reported – referenced as 1977a and 1977b).

Davis, R., & Maxwell, C. (2002) *Preventing repeat incidents of family violence: A reanalysis of data from three field tests,* New York: Vera Institute of Justice.

Davis, R. C., & Taylor, B. G. (1997) 'A proactive response to family violence: The results of a randomized experiment', *Criminology,* 35(2): 307-333.

Davis, R. C., & Medina-Ariza, J. (2001) *Results from an elder abuse prevention experiment in New York City*, Washington, DC: National Institute of Justice.

Davis, R., Weisburd, D., & Hamilton, E. (2007) 'Preventing repeat incidents of family abuse: A randomized field trial of a second responder program in Redlands, CA.', *Journal of Experimental Criminology 2010,* 6: 397-418.

Dunford, F.W. (1990b) 'System-initiated warrants for suspects of misdemeanor domestic assault: a pilot study', *Justice Quarterly,* 7(4): 631-654.

Dunford, F.W., Huizinga, D. and Elliott, D.S. (1990) 'The role of arrest in domestic assault: the Omaha Police Experiment', *Criminology,* 28(2): 183-206.

Dunford, F.W., Osgood, D.W. and Weichelsbaum, H.F. (1982) *National Evaluation of Diversion Projects: Final Report.* Washington, DC: US Department of Justice, National Institute for Juvenile Justice and Delinquency

Prevention. (Four Police RCTs reported: Kansas, Orange, New York City Transit and Memphis).

Earle, H.H. (1973) *Police Recruit Training: Stress v Non-Stress: a revolution in Law Enforcement Career Programs*, Springfield: Charles Thomas.

Eck, J., & Wartell, J. (1998) 'Improving the Management of Rental Properties with Drug Problems: A Randomized Experiment', *Crime Prevention Studies,* 9: 161-185.

Esbenson, F.-A., Peterson, D., Taylor, J. and Osgood, D.W. (2012) 'Gang Resistance Education and Training Program', *Justice Quarterly,* 29(1): 125-151.

**Farrar, W.A. & Ariel, B. (2013) *Self-awareness to Being Watched and Socially Desirable Behavior: A Field Experiment on the Effect of Body-worn Cameras on Police Use of Force*, Washington, DC: Police Foundation.

Gersons, B.P.R, Carlier, I.V.E., Lamberts, R.D. and van der Kolk, B.A. (2000) 'Randomized Clinical Trial of Brief Eclectic Psychotherapy for Police Officers with Posttraumatic Stress Disorder, *Journal of Traumatic Stress,* 13(2): 333-347.

*Giacomantonio, C., Jonathan-Zamir, T., Litmanovitz, Y., Bradford, B., Davies, M., Strang, L., and Sutherland, A. (2016) *College of Policing Stop and Search Training experiment: process evaluation*, Cambridge: RAND Europe.

Giblin, M. J. (2002) 'Using police officers to enhance the supervision of juvenile probationers: An evaluation of the Anchorage CAN program', *Crime & Delinquency,* 48(1): 116-137.

Glick, B., Hamilton, E. and Forst, B. (1986) *Shoplifting: an experiment in lesser crimes and punishments*, Washington, DC: Police Foundation.

Graziano, L.M., Rosenbaum, D.P. and Schuck, A.M. (2014) 'Building group capacity for problem-solving and police-community partnerships through survey feedback and training: a randomised control trial within Chicago's community policing program', *Journal of Experimental Criminology,* 10(1): 79-103.

Groff, E, Kearley, B., Foog, H., Beatty, P., Couture, H. and Wartell, J. (2005) 'A randomized experimental study of sharing crime data with citizens: Do maps produce more fear?', *Journal of Experimental Criminology,* 1: 87-115.

Groff, E., Ratcliffe, J., Haberman, C., Sorg, E., Joyce, N. and Taylor, R.B. (2015) 'Philadelphia Police Tactics Experiment', *Criminology*, 53(1): 23-52.

Grossmith, L., Owens, C., Finn, W., Mann, D., Davies, T. and Baika, L. (2015) *Police, Camera, Evidence: London's Cluster Randomised Controlled Trial of Body Worn Video*, London: National College of Policing and Mayor's Office for Policing and Crime. Retrieved 7th December 2016 from http://whatworks.college.police.uk/Research/Documents/Police_Camera_Evidence.pdf.

Hegarty, T., Williams, L. S., Stanton, S., & Chernoff, W. (2014) 'Evidence-Based Policing at Work in Smaller Jurisdictions', *Translational Criminology,* Spring, 6: 14-15 & 18.

*&**Henstock, D. and Ariel, B. (2017) 'Testing the effects of police body-worn cameras on use of force during arrests: a randomised controlled trial in a large British police force'. *European Journal of Criminology,* Online First: DOI: 10.1177/1477370816686120, 1-31.

Hirschel, D., Hutchison, I. W., Dean, C. W., Kelley, J. J., & Pesackis, C. E. (1990) *Charlotte Spouse Assault Replication Project: Final report*, Charlotte, NC: University of North Carolina at Charlotte.

Ireland, M., Malouff, J.M. and Byrne, B. (2007) 'The efficacy of written emotional expression in the reduction of psychological distress in police officers', *International Journal of Police Science and Management,* 9(4): 303-311.

Jennings, W.G., Lynch, M., & Fridell, L.A. (2015) 'Evaluating the impact of police officer body-worn cameras (BWCs) on response-to-resistance and serious external complaints: Evidence from the Orlando Police Department (OPD) experience utilizing a randomized controlled experiment', *Journal of Criminal Justice,* 43(6): 480-486.

Jolin, A., Feyerherm, W., Fountain, R. and Friedman, S. (1998) *Beyond arrest: The Portland, Oregon domestic violence experiment*, Washington, DC: National Institute of Justice.

**Klein, M.W. (1986) 'Labelling theory and delinquency policy: An experimental test', *Criminal Justice and Behavior,* 13(1): 47-79.

Komro, K.A., Perry, C., Veblen-Mortenson, S., Stigler, M., Munson, K. and Farbakhsh. K. (2004) 'Violence related outcomes of the DARE Plus Project', *Health Education and Behavior,* 31: 335-354.

Koper, C, Taylor, B. and Woods, D.J. (2013) 'A randomised test of initial and residual deterrence from directed patrols and use of license plate readers at crime hot spots', *Journal of Experimental Criminology,* 9: 213-244.

Ku, R. and Blew, C.H. (1977) *The Adolescent Diversion Project in Urbana and Champlain Illinois,* National Institute of Law Enforcement and Criminal Justice. Retrieved 6th December 2016 from https://www.ncjrs.gov/pdffiles1/Digitization/38510NCJRS.pdf.

La Vigne, N. and Lowry, S.S. (2011) *Evaluation of Camera Use to Prevent Crime in Commuter Parking Facilities*, Urban Institute: Justice Policy Center.

Langley, B. (2014) ‘A randomised control trial comparing the effects of procedural justice to experienced utility theories in airport security stops’, [Unpublished M.St. Thesis, Institute of Criminology, University of Cambridge].

**Lincoln, S.B., Teilmann, K.S., Klein, M.W. and Labin, S. (1977) ‘California Youth Authority Diversion Trial’, [Paper presented at the National Conference on Criminal Justice Evaluation, Washington, DC, February, 1977].

*&**Litmanovitz, Y.D. and Montgomery, P. (2015) ‘Procedural Justice Training in the Israeli Police’, *European Journal of Policing Studies,* 3(2), 184-203.

*&**Litmanovitz, Y.D. (2016) ‘Moving towards an evidence-base of democratic police training: The development and evaluation of a complex social intervention in the Israeli Border Police’ [Unpublished D. Phil. Thesis, Centre for Evidence-based Interventions, University of Oxford, Oxford, UK].

Little, M., Kogan, J., Bullock, R. and Van der Laan, P. (2004) ‘An Experiment in Multi-systemic responses to Persistent Young Offenders to Children's Services’, *British Journal of Criminology,* 44: 225-240.

Lu, Fangwen, Zhang, J. and Perloff, J.M. (2012) *Deterring Traffic Violations: Evidence from a Randomised Experiment,* Report for the Natural Science Foundation of China. Retrieved 7th December 2016 from http://www.webmeets.com/files/papers/res/2013/139/Deterrence.pdf.

Lum, C., Hibdon, J., Cave, B., Koper, C., and Merola, L. (2011) ‘License Plate Reader (LPR) Police Patrols in crime hot spots: an experimental evaluation in two adjacent districts’, *Journal of Experimental Criminology,* 7: 321-345.

Lurigio, A.J. and Rosenbaum, D.P. (1992) ‘The travails of the Detroit police-victims experiment: Assumptions and important lessons’, *American Journal of Police,* 11: 1-34.

Martin, S. and Sherman, L.W. (1986) ‘Selective apprehension: A police strategy for repeat offenders’, *Criminology,* 24(1): 155-172.

**MacQueen, S. and Bradford, B. (2015) ‘Enhancing Public Trust and Police Legitimacy during road traffic encounters: results from a randomised controlled trial in Scotland’, *Journal of Experimental Criminology,* 11(3): 419-443.

McCold, P., and Wachtel, B. (1998) *The Bethlehem Pennsylvania police family group conferencing project- violence experiment,* Pipersville: Pipers Press.

McCraty, R. and Tomasino, D. (1999) ‘Resilience Training Program Reduces Physiological and Psychological Stress in Police Officers’, *Global Advances in Health and Medicine,* 1(5): 42-64.

McEwen, J.T., Connors, E.F. and Cohen, M.I. (1986) *Differential Police Response Field Test,* Washington, DC: National Institute of Justice. (Three RCTs in Garden Grove (CA), Greensboro (NC) and Toledo (OH)).

McGarrell, E. and Kroovand, N. (2007) ‘Family Group Conferencing and re-offending amongst First Time Juvenile Offenders: the Indianapolis Experiment’, *Justice Quarterly,* 24(2): 221-246

Mazerolle, L., Bennett, S., Antrobus, E. and Eggins, E. (2012) ‘The Queensland Community Engagement Trial (QCET)’, *Journal of Experimental Criminology,* 8: 343-367.

**Mazerolle, L., Kadleck, C. and Roehl, J. (1988) ‘Controlling drug and disorder problems: the role of place managers’, *Criminology,* 36: 371-404.

**Mazerolle L. G., Price J. F., & Roehl J. (1990). ‘Civil remedies and drug control: a randomized field trial in Oakland, CA.’, *Evaluation Review,* 24(2): 212-241.

Mejia, D., Ortega, D. and Garcia, J.F. (2013) *Police Reform, Training and Crime: Experimental evidence from Columbia's Plan Cuadrantes*, Fundacion Ideas para la Paz. Retrieved 8th December 2016 from http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?fecvnodeid=127138&groupot593=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&dom=1&fecvid=21&ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&v21=127138&lng=en&id=15976.

Mohler, G.O., Short, M.B., Malinowski, S., Johnson, M., Tita, G.E., Bertozzi, A.L. and Brantingham, P.J. (2015) 'Randomized Controlled Field Trials of Predictive Policing', *Journal of the American Statistical Association,* 110(512): 1399-1411. (2 RCTs are mentioned but only one – LAPD - clearly completed)

Neyroud, P.W. and Slothower, M., Sherman, L.W., Neyroud, E.C. and Ariel, B. (2015) 'Operation Turning Point: Second Interim Report', [Unpublished report, West Midlands Police and Institute of Criminology, University of Cambridge].

Norvell, N. and Belles, D. (1993) 'Psychological and Physical Benefits of Circuit Weight Training in Law Enforcement Personnel', *Journal of Consulting and Clinical Psychology,* 61(3): 520-527.

Owens, C., Mann, D. and McKenna, R. (2014) *The Essex Body Worn Video Trial,* London: National College of Policing. Retrieved 7th December 2016 from http://whatworks.college.police.uk/Research/Documents/BWV_Report.pdf.

Owens, E., Weisburd, D.W., Alpert, G. and Amendola, K. (2015) *Promoting Officer Integrity through Early Engagements and Procedural Justice in the Seattle Police Department*, Washington, DC: Police Foundation. Retrieved 8th December 2016 from https://www.ncjrs.gov/pdffiles1/nij/grants/249881.pdf.

Pate, A. M, Lavrakas, P. J., Wycoff, M. A., Skogan, W. G., & Sherman, L. W. (1985) *Neighborhood Police Newsletters: Experiments in Houston, Technical Report,* Washington, DC: Police Foundation.

Pate, A. M, Lavrakas, P. J., Wycoff, M. A., Skogan, W. G., & Sherman, L. W. (1985) *Neighborhood Police Newsletters: Experiments in Newark, Technical Report,* Washington, DC: Police Foundation.

Pate, A. M. & Hamilton, E. E. (1992) 'Formal and informal deterrents to domestic violence: The Dade County Spouse Assault Experiment', *American Sociological Review,* 57(5): 691-698.

Pate, A. M., Hamilton, E. E. and Sampson Annan (1991) *Safe Streets Experiment - Dade County*, Washington, DC: Police Foundation.

Piza, E.L., Caplan, J.M., Kennedy, L.W. and Gilchrist, A.M. (2015) 'The effects of merging proactive CCTV monitoring with direct police patrol: a randomised controlled trial', *Journal of Experimental Criminology,* 11(1): 43-69.

Quay, H. and Love, C. (1977) 'The Juvenile Services Program, Pinellas County Florida', *Criminal Justice and Behavior,* 4(4): 377-396.

Quinton, P. (2011) *The impact of information about crime and policing on public perceptions: the results of a randomised controlled trial*, London: National College of Policing. Retrieved 8th December 2016 from http://www.college.police.uk/en/docs/Full_Report_-_Crime_and_Policing_Information.pdf.

Ratcliffe, J., Taniguchi, T., Groff, E. R., Wood, J. D. (2011) 'The Philadelphia Foot Patrol Experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots', *Criminology,* 49(3): 795-831.

Ridgeway, G., Braga, A. A., Tita, G., & Pierce, G. L. (2011) 'Intervening in gun markets: An experiment to assess the impact of targeted gun-law messaging', *Journal of Experimental Criminology*, 7(1): 103-109.

Ringwalt, C., Ennett, S. T., & Holt, K. D. (1991) 'An outcome evaluation of Project DARE (Drug Abuse Resistance Education)', *Health Education Research,* 6(3): 327-337.

Roman, J.K., Reid, S.E., Chalfin, A.J. and Knight, C.R. (2009) 'The DNA Field Experiment: a randomised controlled trial of using DNA to solve property crimes', *Journal of Experimental Criminology,* 5: 345-369.

Rose, G. and Hamilton, R. (1970) 'The Effects of a Juvenile Liaison Scheme*'* *British Journal of Criminology,* 10(1): 2-20.

Rosenbaum, D.P., Lurigio, A.J. and Lavrakas, P.J. (1989) *A National Evaluation of the Crimestoppers Programme*, Washington, DC: Police Foundation.

Rosenbaum, D. P., Flewelling, R. L., Bailey, S. L., Ringwalt, C. L., & Wilkinson, D. L. (1994) 'Cops in the Classroom: A Longitudinal Evaluation of Drug Abuse Resistance Education (DARE)', *Journal of Research in Crime & Delinquency,* 31(1): 3-31.

Rosenbaum, D.P. and Lawrence, D.S. (2013) *Teaching Respectful Police-Citizen Encounters and Good Decision Making: Results of a Randomized Control Trial with Police Recruits*, Retrieved 8th December 2016 from http://static1.1.sqspcdn.com/static/f/733761/16725011/1329855686737/Rosenbaum++LawrenceTeaching+Respectful+Police-Cit++Encounters.pdf?token=3B4ojj6Hig8Bvc9FVWLiNFOqP1E%3D.

Rosenfeld, R., Deckard, M.J. and Blackburn, E. (2014) 'The Effects of directed patrol and self-initiated enforcement on firearm violence', *Criminology,* 52(3): 428-449.

Sahin, Nusret, M. (2014) ‘Legitimacy, Procedural Justice and Police-Citizen Encounters: A Randomized Controlled Trial of the Impact of Procedural Justice on Citizen Perceptions of the Police during Traffic Stops in Turkey’, [Unpublished PhD. Thesis, Rutgers University].

Santos, R.B. and Santos, R.G. (2014) *An experimental test of offender-based strategies in residential burglary and theft from vehicle hot spots*, Washington DC: Bureau of Justice Assistance, Smart Policing Initiative.

*Santos, R.B. and Santos, R.G. (2016) ‘Offender focused police intervention in residential burglary and theft from vehicle hotspots: a partially blocked randomized trial.’, *Journal of Experimental Criminology,* 12(3): 373-402.

**Shapland, J., Atkinson, A., Atkinson, H., Chapman, B., Colledge, E., Dignan, J., Howes, M., Johnstone, J., Robinson, G. and Sorsby, A. (2006) *Restorative Justice in Practice: the second report from the evaluation of three schemes,* London: Home Office. (Six Restorative Justice Police RCTs: London - Robbery over 18 and Burglary over 18; Northumbria – Violence over 18, Property Crimes over 18, Violence under 18 and Property Crimes under 18 – see also Sherman et al. (2015) below).

Sherman, L.W. and Berk, R.A. (1984) ‘The Specific Deterrent Effects of Arrest for Domestic Assault’, *American Sociological Review,* 49(2): 261-272.

**Sherman, L.W., Buerger, M.E. and Gartin, P.R. (1989) *Repeat Call Address Policing: The Minneapolis RECAP Experiment. Final Report*, Washington, DC: National Institute of Justice and Crime Control Institute.

Sherman, L. W., & Rogan, D. P. (1995) ‘Deterrent effects of police raids on crack houses: A randomized, controlled experiment’ *Justice Quarterly,* 12(4): 755-781.

**Sherman, L.W., Strang, H.S., Barnes, G., Woods, D.J., Bennett, S., Inkpen, N., Newbury-Birch, D., Rossner, M., Angel, C., Mearns, M. and Slothower, M.P. (2015) 'Twelve Experiments in Restorative Justice: The Jerry Lee program of randomized trials of restorative justice conferences', *Journal of Experimental Criminology,* 11(4): 501-540.

Sherman, L. W., Schmidt, J. D., Rogan, D. P., Gartin, P. R., Cohn, E. G, Collins, D. J., & Bacich, A. R. (1996) 'The variable effects of arrest on criminal careers: The Milwaukee Domestic Violence Experiment', *Journal of Criminal Law and Criminology,* 83(1): 137-169.

Sherman, L.W. and Weisburd, D.W. (1995) 'General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial', *Justice Quarterly,* 12(4): 625-648.

Shipley, P. and Baranski, J. (2002) 'Police Office Performance under Stress: a pilot study on the effects of visuo-motor behavior rehearsal', *International Journal of Stress Management,* 9(2): 71-80.

Short, M.A., DiCarlo, S., Steffee, W.P. and Pavlou, K. (1984) 'Effects of Physical Conditioning on Self-Concept of Adult Obese Males', *Physical Therapy,* 64(2): 194-198.

Skogan, W. and Wycoff, M.A. (1987) 'Some unexpected effects of a police service for victims', *Crime and Delinquency,* 33: 490-501.

Sloboda, Z., Stephens, R. C., Stephens, P. C., Grey, S. F., Teasdale, B., Hawthorne, R. D., Williams, J., & Marquette, J. F. (2009) 'The Adolescent Substance Abuse Prevention Study: A randomized field a universal substance abuse prevention program', *Drug and Alcohol Dependence,* 102: 1-10.

Slothower, M. (2014a). 'Victim Satisfaction and Perceptions of Police and Criminal Justice Legitimacy: Mediating Impacts of Perceptions of Problem-

Solving Motive Achievement and Matching', [Unpublished M.Phil. Thesis, University of Maryland].

Sousa, W., Ready, J. and Ault, M. (2010) 'The Impact of TASER's on police use of force decisions: findings from a randomised field-training experiment', *Journal of Experimental Criminology,* 6: 35-55

**Strang, H., Barnes, G.C., Braithwaite, J. and Sherman, L.W. (1999). *Experiments in Restorative Policing: a Progress Report on the Canberra Reintegrative Shaming Experiments (RISE)*, Retrieved 8th December 2016 from http://www.aic.gov.au/media_library/aic/rjustice/rise/progress/1999.pdf.

Stratton, J.G. (1975) 'Effects of Crisis Intervention Counseling on Predelinquent and Misdemeanor Juvenile Offenders*' Juvenile Justice,* 26: 7-18.

Tanigoshi, H., Kontos, A.P., and Remley, T.P. (2008) 'The effectiveness of individual wellness counseling on the wellness of Law Enforcement Officers', *Journal of Counseling and Development,* 86(1): 64-86.

Taylor, B., Davis, R.C. and Maxwell, C.D. (2001) 'The effect of a batterer treatment program: a randomised experiment in Brooklyn', *Justice Quarterly,* 18: 171-201.

Taylor, B., Koper, C. S., Woods, D. J. (2011) 'A randomized controlled trial of different policing strategies at hot spots of violent crime', *Journal of Experimental Criminology,* 7(2): 149-181.

Telep, C., Mitchell, R. and Weisburd, D.W. (2014) 'How Much Time Should the Police Spend at Crime Hot Spots? Answers from a Police Agency Directed Randomized Field Trial in Sacramento, California', *Justice Quarterly,* 31(5): 905-933.

Weisburd, D. & Green L. (1995) 'Policing drug hot spots: The Jersey City Drug Market Analysis Experiment', *Justice Quarterly,* 12: 711-736.

Weisburd, D.W., Groff, E., Jones, G., Cave, B., Amendola, K., Yang, S-M. and Emison, R. (2015). 'The Dallas Patrol Management Experiment: Can AVL Technologies be Used to Harness Unallocated Patrol Time for Crime Prevention?', *Journal of Experimental Criminology,* 11(3): 367-391.

Weisburd, D., Hinkle, J.C., Famega, C. and Ready, J. (2011) 'The possible "backfire" effects of hot spots policing: an experimental assessment of impacts of legitimacy, fear and collective efficacy', *Journal of Experimental Criminology*, 7(2): 297-320.

Weisburd, D., Morris, N., & Ready, J. (2008) 'Risk-focused policing at places: An experimental evaluation', *Justice Quarterly* 25(1): 163-200.

Wells, W., Horney, J., Maguire, E. (2005) 'Patrol Officers responses to citizen feedback: an experimental analysis', *Police Quarterly,* 8: 171-205.

**Wells, G, Steblay, N. and Dysart, J. (2011) *A test of the simultaneous vs. sequential line-up methods: an initial report of the AJS national eyewitness identification field studies*. Des Moines, Iowa: American Judicature Society.

**Wells, G., Steblay, N. and Dysart, J. (2015) 'A test of the simultaneous vs. sequential line-up methods: an initial report of the AJS national eyewitness identification field studies', *Law and Human Behavior,* 39(1): 1–14.

Wheller, L., Quinton, P., Fildes, A., and Mills, A. (2013) *The Greater Manchester Procedural Justice Experiment: the impact of communication skills training on officers and victims of crime*. London: National College of Policing. Retrieved 8th December 2016 from http://library.college.police.uk/docs/college-of-policing/Practitioner-Paper.pdf.

Wilson, S.A., Tinker, R.H., Becker, L.A. and Logan, C.R. (2001) ‘Stress Management with Law Enforcement Personnel: a controlled outcome study of EMDR versus a traditional Stress Management Program’, *International Journal of Stress Management,* 8(1): 179-200.

APPENDIX 1: Randomised Controlled Trials in Policing 1970-2

Author(s)	Title	Coded RCT topic ((after Braga et al., 2014)	Topic Code Number	SAMPLE	Unit	Treat as Assigned/treatment integrity %	Novice Investigator team	New Research station	Novel topic or intervention	Practitioner involved
Abrahamse et al. (1991)	An experimental evaluation of the Phoenix repeat offender program	offender management	7	480	Repeat Offenders Repeat Offenders	91.25	U	Y	N	N
Ackerley (1986)	The Effects of a stress management program on Police Personnel	Health and Welfare	13	49	Police Officers	Unclear	Y	Y	Y	N
Amendola et al. (2011)	An experimental study of compressed work schedules in policing: advantages and disadvantages of various shift lengths	Health and Welfare	13	326	Police Officers	84.35	N	Y	Y	N
Angel (2014)	Crime victims meet their offenders: testing the impact of restorative justice conferences on victims' post-traumatic stress symptoms	Restorative Justice	2	192	Victims	99.47	N	Y	N	N
Antrobus, E. and Pilotto, A. (2016)	The Queensland Unlawful Entry Experiment	Crime Victim outreach	9	978	Crime Scenes	93.5	N	N	Y	Y
Ariel, et al. (2014)	The Effect of Police Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police:A Randomized Controlled Trial	Body Worn Cameras	15	988	Shifts/Tours of duty	Unclear	Y	Y	Y	Y

Banerjee et al. (2013)	Randomised Drunk Driving Crackdown in India	Road Safety - Drink Drive	16	123	Police Stations and Road Breath Stations	53	Y	Y	Y	Y
Banerjee et al. (2012)	Can Institutions be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police	Adminstrative Reforms	17	162	Police Stations	84	Y	Y	Y	Y
Bennett and Newman (2015)	Mobile Community Police Office Experiment (Queensland)	Fear Reduction	8	26	hotspots in N Brisbane	Unclear	N	N	Y	Y
Berk et al. (1992)	Colorado Springs Spouse Abuse Experiment	Domestic violence	3	1658	DV suspects	82	N	Y	N	N
Binder and Newkirk (1977)	University of California Youth Services Program	Juvenile justice	5	No details provided	juveniles	Unclear	Y	Y	Y	N
Boyanowsky and Griffiths (1982)	Weapons and Eye Contact as Instigators or Inhibitors of Aggressive Arousal in Police citizen Interaction	Legitimacy	18	133	Citizens stopped at roadside	Unclear	Y	Y	Y	N
Braga and Bond (2008)	Policing crime and disorder hot spots: A randomized, controlled trial	Hotspots/crime places	1	34	Hotspots	Unclear	N	Y	N	N
Braga et al. (1999)	Problem-oriented policing in violent crime places: A randomized controlled experiment	Hotspots/crime places	1	24	Violent Crime Hotspots	Unclear	N	N	Y	Y
Byles and Maurice (1979)	The juvenile services project: an experiment in delinquency control	Juvenile justice	5	305	Juveniles with 2 or more priors	45	Y	Y	Y	N
Clayton et al. (1996)	The effectiveness of Drug Abuse Resistance Education (Project DARE): 5-year follow-up results	DARE	4	31	Schools and Pupils in 6th Grade	93	Y	Y	N	N

Davidson et al. (1977a)	The Urbana-Champaign Pre-court diversion trial (1)	Juvenile justice	5	37	Juvenile offenders	Unclear	Y	Y	Y	N
Davidson et al. (1977b)	The Urbana-Champaign Pre-court diversion trial (2)	Juvenile justice	5	36	Juvenile offenders	Unclear	N	N	N	N
Davis and Taylor (1997)	A proactive response to family violence: The results of a randomized experiment	Domestic violence	3	435	Households in which family violence occurred	84	Y	Y	Y	N
Davis and Medina-Ariza (2001)	Results from an elder abuse prevention experiment in New York City	Domestic violence	3	60	Housing Projects/Households	50	N	Y	Y	N
Davis and Maxwell (2002)	Preventing repeat incidents of family violence: A reanalysis of data from three field tests	Domestic violence	3	197	victims of family violence	Unclear	N	N	Y	N
Davis et al. (2007)	Preventing repeat incidents of family abuse: A randomized field trial of a second responder program in Redlands, CA	Domestic violence	3	300	Victims of Domestic Violence	85	N	Y	N	N
Dunford et al. (1990)	The role of arrest in domestic assault: the Omaha Police Experiment	Domestic violence	3	330	Offenders in DV cases	92	N	Y	N	N
Dunford (1990)	System-initiated warrants for suspects of misdemeanor domestic assault: A pilot study	Domestic violence	3	247	Offenders in DV cases	96.5	N	Y	Y	N
Dunford et al. (1982)	National Evaluation of Diversion Pilots: **Kansas**	Juvenile justice	5	433	Juvenile offenders	84.7	Y	Y	Y	N
Dunford et al. (1982)	National Evaluation of Diversion Pilots: Orange	Juvenile justice	5	686	Juvenile offenders	69.5	Y	Y	Y	N
Dunford et al. (1982)	National Evaluation of Diversion Pilots: New York City Transit Police	Juvenile justice	5	533	Juvenile offenders	78.2	Y	Y	Y	N

Dunford et al. (1982)	National Evaluation of Diversion Pilots: Memphis	Juvenile justice	5	975	Juvenile offenders	54.35	Y	Y	Y	N
Earle (1973)	Police Recruit Training: Stress v Non-Stress: a field experiment in LA Sheriffs Department	Police Training	19	174	Deputy Sheriffs	57.5	Y	Y	Y	Y
Eck and Wartell (1998)	Improving the Management of Rental Properties with Drug Problems: A Randomized Experiment	Hotspots/crime places	1	121	Residential properties subject of drugs enforcemem t	83	Y	Y	Y	N
Esbenson et al. (2012)	Gang Resistance Education and Training Program	DARE	4	195	Classrooms	90	N	Y	N	N
Gersons et al. (2000)	Randomized Clinical Trial of Brief Eclectic Psychotherapy for Police Officers with Posttraumatic Stress Disorder	Health and Welfare	13	42	Police Officers	97.6	N	Y	Y	N
Giblin (2002)	Using police officers to enhance the supervision of juvenile probationers: An evaluation of the Anchorage CAN program	Juvenile justice	5	190	Juvenile offenders	45	Y	Y	Y	N
Glick et al. (1986)	Shoplifting: an experiment in lesser crimes and punishments	shoplifting	20	1346	Shoplifters	92	Y	Y	Y	N
Graziano et al. (2014)	Building group capacity for problem-solving and police-community partnerships through survey feedback and training: a randomised control trial within Chicago's community policing program	Citizen feedback/interaction	10	51	Beats in Chicago	47.5	N	N	Y	N

Groff et al. (2005)	A randomized experimental study of sharing crime data with citizens: Do maps produce more fear?	Citizen feedback/interaction	10	314	Residents	100	N	Y	Y	N
Groff et al. (2015)	Philadelphia Police Tactics Experiment	Hotspots/crime places	1	81	Micro places	Unclear	N	N	Y	N
Grossmith et al. (2015)	Police, Camera, Evidence: London's Cluster Randomised Controlled Trial of Body Worn Video.	Body Worn Cameras	15	1510	Officers	Unclear	N	Y	N	Y
Hegarty et al. (2014)	Evidence-Based Policing at Work in Smaller Jurisdictions	Hotspots/crime places	1	48	Hotspots of crime	88	Y	Y	N	Y
Hirschel et al. (1990)	Charlotte Spouse Assault Replication Project: Final report	Domestic violence	3	686	Cases of Domestic Violence	83.5	Y	Y	N	N
Ireland et al. (2007)	The efficacy of written emotional expression in the reduction of psychological distress in police officers	Health and Welfare	13	129	Police Officers	51.9	Y	Y	Y	N
Jennings et al. (2015)	Orlando Police Department Body Worn Camera Experiment	Body Worn Cameras	15	89	Police Officers	Unclear	Y	Y	N	N
Jolin et al. (19	Beyond arrest: The Portland, Oregon domestic violence experiment	Domestic violence	3	927	DV Incidents	70.7	Y	Y	Y	N
Klein (1986)	Labeling theory and delinquency policy: An experimental test	Juvenile justice	5	306	Juvenile offenders	Unclear	Y	Y	Y	N
Komro et al. (2004)	Violence related outcomes of the DARE Plus Project	DARE	4	24	Schools and 7th Grade pupils	98	N	Y	N	N

Koper et al. (2013)	A randomised test of initial and residual deterrence from directed patrols and use of license plate readers at crime hot spots	Hotspots/crime places	1	117	Hot route sites	100	N	N	N	N
Ku and Blew (1977)	The Adolescent Diversion Project in Urbana and Champlain Illinois	Juvenile justice	5	36	adolescent offenders	100	Y	Y	Y	N
La Vigne and Lowry (2011)	Evaluation of Camera Use to Prevent Crime in Commuter Parking Facilities	CCTV	21	50	Parking facilities	100	Y	Y	Y	N
Langley (2014)	A randomised control trial comparing the effects of procedural justice to experienced utility theories in airport security stops	Legitimacy	18	781	Passengers	98.4	Y	N	Y	Y
Lincoln et al. (1977)	California Youth Authority Diversion Trial	Juvenile justice	5	306	Juvenile offenders	Unclear	Y	Y	Y	N
Little et al. (2004)	An Experiment in Multi-systemic responses to Persistent Young Offenders to Children's Services	Juvenile justice	5	90	Juvenile offenders	64	Y	Y	Y	N
Lu et al. (2012)	Deterring Traffic Violations: Evidence from a Randomised Experiment	Road Safety: traffic enforcement	16	80377	Private car owners	100	Y	Y	Y	N
Lum et al. (2011)	License Plate Reader (LPR) Police Patrols in crime hot spots: an experimental evaluation in two adjacent districts	Hotspots/crime places	1	30	Hotspots	84.6	N	Y	Y	N
Lurigio and Rosenbaum (1992)	The travails of the Detroit police-victims experiment: Assumptions and important lessons	Crime Victim outreach	9	122	Police Recruits	26.5	N	Y	Y	N

MacQueen and Bradford (2014)	ScotCet - Scottish Commmunity Engagement Trial	Legitimacy	18	20	Police Units conducting Road Safety tests	Unclear	Y	Y	N	N
Martin and Sherman (1986)	Selective apprehension: A police strategy for repeat offenders	Offender management	7	414	Repeat Offenders	80	N	Y	Y	N
McCold and Wachtel (1998)	The Bethlehem Pennsylvania police family group conferencing project-violence experiment	Restorative Justice	2	111	juvenile offenders	31.6	Y	Y	Y	N
McCold and Wachtel (1998)	The Bethlehem Pennsylvania police family group conferencing project property experiment	Restorative Justice	2	181	juvenile offenders	48.6	Y	Y	Y	N
McCraty and Tomasino (1999)	Resilience Training Program Reduces Physiological and Psychological Stress in Police Officers	Health and Welfare	13	65	Police Officers	90.7	Y	Y	Y	N
McEwen et al. (1986)	Differential Police Response Field Test - Garden Grove, California	Differential Police Response	25	5510	Calls from public	90.7	Y	Y	Y	N
McEwen et al. (1986)	Differential Police Response Field Test	Differential Police Response	25	34795	Calls from public	91.3	Y	Y	Y	N
McEwen et al. (1986)	Differential Police Response Field Test - Toldeo	Differential Police Response	25	5497	Calls from public	98	Y	Y	Y	N
McGarrell and Kroovand (2007)	Family Group Conferencing and re-offending amongst First Time Juvenile Offenders: the Indianapolis Experiment	Juvenile justice	5	782	Juvenile offenders	98.5	Y	Y	N	N
Mazerolle et al. (2012)	The Queensland Community Engagement Trial (QCET).	Legitimacy	18	60	Random Breath Test Stations	100	N	Y	Y	N

Mazerolle et al. (1998)	Controlling drug and disorder problems: the role of place managers	Hotspots/crime places	1	100	Street Blocks	70	Y	Y	Y	N
Mazerolle et al. (2000)	Civil remedies and drug control: a randomized field trial in Oakland, CA	Hotspots/crime places	1	100	Street Blocks	70	Y	Y	Y	N
Mejia et al. (2013)	Police Reform, Training and Crime: Experimental evidence from Columbia's Plan Cuadrantes	Police Training	19	104	Police stations	Unclear	Y	Y	Y	N
Mohler et al. (2015)	Randomized Controlled Field Trials of Predictive Policing (Los Angeles)	Predictive Policing	22	510	Patrol days	Unclear	Y	Y	Y	N
Neyroud et al. (2015)	Operation Turning Point: a randomised controlled trial of offender desistance policing	Pre-court diversion	6	417	Offenders	91	N	N	Y	Y
Norvell and Belles (1993)	Psychological and Physical Benefits of Circuit Weight Training in Law Enforcement Personnel	Health and Welfare	13	48	Police Officers	64.4	Y	Y	Y	N
Owens et al. (2014)	The Essex Body Worn Video Trial	Body Worn Cameras	15	300	Police Officers	Unclear	Y	Y	N	Y
Owens et al. (2015)	Promoting Officer Integrity through Early Engagements and Procedural Justice in the Seattle Police Department.	Legitimacy	18	1444	Police Officers	100	N	N	Y	N
Pate et al. (1985a)	Neighborhood Police Newsletters: Experiments in Houston, Technical Report	Citizen feedback/interaction	10	660	Households	Unclear	N	Y	Y	N
Pate et al. (1985b)	Neighborhood Police Newsletters: Experiments in Newark, Technical Report	Citizen feedback/interaction	10	504	Households	Unclear	N	Y	Y	N

Pate and Hamilton (1992)	Formal and informal deterrents to domestic violence: The Dade County Spouse Assault Experiment	Domestic violence	3	907	Cases of Domestic Violence	89.9	N	Y	N	N
Pate et al. (1991)	Safe Streets Experiment - Dade County	Domestic violence	3	907	Victims of Domestic Violemce	61.3	N	Y	Y	N
Piza et al. (2015)	The effects of merging proactive CCTV monitoring with direct police patrol: a randomised controlled trial	CCTV	21	38	CCTV Camera areas	Unclear	Y	N	Y	N
Quay and Love (1977)	The Juvenile Services Program, Pinellas County Florida	Juvenile justice	5	568	Juvenile offenders	Unclear	Y	Y	Y	N
Quinton (2011)	The impact of information about crime and policing on public perceptions: the results of a randomised controlled trial	Citizen feedback/interaction	10	7434	Citizens	100	Y	Y	Y	Y
Ratcliffe et al. (2011)	The Philadelphia Foot Patrol Experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots	Hotspots/crime places	1	60	Violent Crime Hotspots	Unclear	N	Y	N	N
Ridgeway et al. (2011)	Intervening in gun markets: An experiment to assess the impact of targeted gun-law messaging	Gun possession	23	2120	Gun purchasers	68.2	N	Y	Y	N
Ringwalt et al. (1991)	An outcome evaluation of Project DARE (Drug Abuse Resistance Education)	DARE	4	20	Schools and pupils	100	Y	Y	Y	N
Roman et al. (2009)	The DNA Field Experiment: a randomised field trial of the cost-effectiveness of using DNA to solve property crimes	DNA/Crime detection	11	2150	Volume crime cases	99.9	Y	Y	Y	N

Rose and Hamilton (1970)	Effects of a juvenile liaison scheme	Juvenile justice	5	494	Young people	Unclear	Y	Y	Y	N
Rosenbaum et al. (1989)	A National Evaluation of the Crimestoppers Programme	Crimestoppers	24	44	Anonymous Callers to Crimestoppers	80	N	Y	Y	N
Rosenbaum et al. (1994)	Cops in the Classroom: A Longitudinal Evaluation of Drug Abuse Resistance Education (DARE)	DARE	4	24	Schools and pupils	Unclear	N	Y	Y	N
Rosenbaum and Lawrence (2013)	Teaching Respectful Police-Citizen Encounters and Good Decision Making: Results of a Randomized Control Trial with Police Recruits	Legitimacy	18	157	Police Recruits	Unclear	N	N	Y	N
Rosenfeld et al. (2014)	The Effects of directed patrol and self initiated enforcement on firearm violence	Hotspots/crime places	1	32	Violent/Firearms Crime Hotspots	Unclear	Y	Y	Y	N
Sahin (2014)	Legitimacy, Procedural Justice and Police-Citizen Encounters: A Randomized Controlled Trial of the Impact of Procedural Justice on Citizen Perceptions of the Police during Traffic Stops in Turkey	Legitimacy	18	702	Drivers stoppped in speed checks	96.6	Y	Y	N	N
Santos and Santos (2014)	An experimental test of offender-based strategies in residential burglary and theft from vehicle hot spots	Hotspots/crime places	1	48	Hotspots	95.9	Y	Y	Y	Y
Shapland et al. (2006)	London RJ RCT -Robbery Offences: Offenders over 18	Restorative Justice	2	106	Offenders	92.2	N	Y	N	N

Shapland et al. (2006)	London RJ RCT - Burglary Offences	Restorative Justice	2	186	Offenders	90.3	N	Y	N	N
Shapland et al. (2006)	Northumbria RJ RCT - Violence Offences - Under 18	Restorative Justice	2	165	Offenders	95.1	N	Y	N	N
Shapland et al. (2006)	Northumbria RJ RCT - Property Offences - Under 18	Restorative Justice	2	165	Offenders	98.2	N	Y	N	N
Shapland et al. (2006)	Northumbria RJ RCT - Property Offences - Over 18	Restorative Justice	2	105	Offenders	95.7	N	Y	N	N
Shapland et al. (2006)	Northumbria RJ RCT - Violence Offences - Over 18	Restorative Justice	2	105	Offenders	93.3	N	Y	N	N
Sherman et al. (1989)	Repeat Call Address Policing: The Minneapolis RECAP Experiment. Final Report	Hotspots/crime places	1	500	Repeat Call addresses	Unclear	N	N	Y	N
Sherman and Weisburd (1995)	General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial	Hotspots/crime places	1	110	Crime Hot spots	91	N	Y	Y	N
Sherman and Berk (1984)	The specific deterrent effects of arrest for domestic assault	Domestic violence	3	314	Domestic Violence cases	83	Y	Y	Y	N
Sherman and Rogan (1995)	Deterrent effects of police raids on crack houses: A randomized, controlled experiment	Hotspots/crime places	1	207	Blocks in Kansas City	97	N	Y	Y	N
Sherman et al. (1992)	The variable effects of arrest on criminal careers: The Milwaukee Domestic Violence Experiment	Domestic violence	3	1200	Domestic Violence cases	98.25	N	Y	N	Y
Short, et al. (1984)	Effects of Physical Conditioning on Self-Concept of Adult Obese Males	Health and Welfare	13	45	Police Officers	100	Y	Y	Y	N

Strang et al. (1999)	EXPERIMENTS IN RESTORATIVE POLICING: A PROGRESS REPORT on the Canberra Reintegrative Shaming Experiments (RISE) - RISE Drink Drive RCT	Restorative Justice	2	900	Offenders	93	N	Y	Y	N
Strang et al. (1999)	RISE Shoplifting RCT	Restorative Justice	2	143	Offenders	88.5	N	Y	Y	N
Strang et al. (1999)	RISE Property Crime RCT	Restorative Justice	2	249	Offenders	76.9	N	Y	Y	N
Strang et al. (1999)	RISE Violent Crime RCT	Restorative Justice	2	121	offenders	86	N	Y	Y	N
Shipley and Baranski (2002)	Police Office Performance under Stress: a pilot study on the effects of visuo-motor behavior rehearsal	Health and Welfare	13	54	Police Officers	100	Y	Y	Y	Y
Skogan and Wycoff (1987)	Some unexpected effects of a police service for victims	Crime Victim outreach	9	485	Victims of personal household crimes	85	Y	Y	Y	N
Sloboda, et al. (2009)	The Adolescent Substance Abuse Prevention Study: A randomized field a universal substance abuse prevention program	DARE	4	83	School clusters	75	Y	Y	N	N
Slothower (2014a)	Operation Turning Point: a randomised controlled trial of victim satisfaction in deferred prosecution and prosecution of the offenders in their crimes	Crime Victim outreach	9	142	victims of crime	91	N	N	Y	Y
Sousa et al. (2010)	The Impact of TASER's on police use of force decisions: findings from a randomised field-training experiment	TASER	12	64	Police Officers	100	N	Y	Y	N

Stratton (1975)	Effects of Crisis Intervention Counseling on Predelinquent and Misdemeanor Juvenile Offenders	Juvenile justice	5	60	juvenile offenders	Unclear	Y	Y	Y	N
Tanigoshi et al. (2008)	The effectiveness of individual wellness counseling on the wellness of Law Enforcement Officers.	Health and Welfare	13	60	Police Officers	85	Y	Y	Y	N
Taylor et al. (2001)	The effect of a batterer treatment program: a randomised experiment in Brooklyn	Domestic violence	3	376	Court Defendants	85	N	Y	Y	N
Taylor et al. (2011)	A randomized controlled trial of different policing strategies at hot spots of violent crime.	Hotspots/crime places	1	83	Violent crime hotspots	Unclear	N	Y	Y	N
Telep et al. (2014)	How Much Time Should the Police Spend at Crime Hot Spots? Answers from a Police Agency Directed Randomized Field Trial in Sacramento, California	Hotspots/crime places	1	42	Hotspots	100	N	Y	Y	Y
Weisburd and Green (1995)	Policing drug hot spots: The Jersey City Drug Market Analysis Experiment	Hotspots/crime places	1	56	Drug Hotspots	100	N	Y	Y	N

D. Weisburd et al. (2015)	The Dallas Patrol Management Experiment: Can AVL Technologies be Used to Harness Unallocated Patrol Time For Crime Prevention?	Hotspots/crime places	1	232	Police beats	100	N	Y	Y	N
Weisburd et al. (2011)	The possible "backfire" effects of hot spots policing: an experimental assessment of impacts of legitimacy, fear and collective efficacy	Hotspots/crime places	1	110	Street Segments	78.57	N	N	N	N
Weisburd et al. (2008)	Risk-focused policing at places: An experimental evaluation	Hotspots/crime places	1	26	Census Blocks	Unclear	N	Y	Y	N
Wells et al. (2005)	Patrol Officers reponses to citizen feedback: an experimental analysis	Citizen feedback/interaction	10	57	Police Officers	84	Y	Y	Y	N
Wells et al. (2015)	A test of the simultaneous vs. sequential lineup methods: an initial report of the AJS national eyewitness identification field studies	Eye-witness identification	14	497	Line-ups	100	Y	Y	Y	N
Wheller et al. (2013)	The Greater Manchester Procedural Justice Experiment: the impact of communication skills training on officers and victims of crime	Police Training	19	576	Police Officers	86.1	N	Y	Y	Y

Wilson et al. (2001)	Stress Management with Law Enforcement Personnel: a controlled outcome study of EMDR versus a traditional Stress Management Program	Health and Welfare	13	62	Police Officers	95	Y	Y	Y	N

APPENDIX 2: IN-FLIGHT POLICE RCTs: Police RCTs completed or in progress, but not published by January 2016

INVESTIGATORS	YEAR	TITLE/SUBJECT	LOCATION	CODED CATEGORY	PRACTITIONER	STATUS
Strang, H., Chilton, S., Cornelius, N. and Braddock, B.	2012	Operation CARA:	Hampshire UK	Domestic violence	Y	Data gathering continuing - interim results reported at Cambridge EBP Conference 2016
Ariel, B., Strang, H. and Sherman, L.W.	2010	Repeat calls for Anti-Social Behaviour	Birmingham UK	Problem-oriented Policing	Y	Results differed between two sites. Data currently being analysed.
Ariel, B. and Newton, M.	2011	Operation Beck	British Transport Police	Hotspots	Y	Study complete - publication awaited. Results presented at Cambridge EBP Conference 2015
Mazerolle, L. Bennett, S., Antrobus, E., Bedford, L., Eggin, E., Prguda, E., Nietzche, F. and Brown-Kenyon, D.	2012	Operation ABILITY	Brisbane, Queensland	Third Party Policing	N	Report in press

Ariel, B. and Smallwood, J. With Lawrence Sherman, Neil Wain, Cristobal Weinborn, Wendy Goodhill, Gabi Sosinski, Justice	2012	Operation Savvy	Birmingham, UK	Hotspots	Y	Study in progress: early results reported at Society of Evidence Based Policing Conference in Birmingham in September 2013 suggest that crime in the treatment areas has reduced significantly
Sherman, L.W., Rose, L., Slothower, M., Wain, N. and Ariel, B.	2013	Trindad and Tobago Hotspot Experiment	Trinidad and Tobago	Hotspots	Y	Study in progress - initial data gathering completed and wider roll out of the strategy under consideration
Linton, B., Herbert, T. and Ariel, B.	2014	Operation MENAS	London UK	Hotspots	Y	Field work complete
Ariel, B., Weinhorn, C. and Sherman, L.W.	2014	Peterborough Hotspots study	Peterborough UK	Hotspots	N	Study complete and report being compiled: results being reported to Cambridge EBP Conference 2014
Kronick, D. and Ortega, D.	2013 to 2015	Sucre Hotspots study	Sucre, Venezuala: www.socialsciencere gistry.org/trials/469/hi story/2336	Hotspots	N	Study in progress in 2014: signficant problems with early implementation and delivery of additional patrol time
Henstock, D. and Ariel, B.	2015	West Midlands Body Worn Video RCT	West Midlands Police (WMP)	Body Worn Video	Y	Article published - Ariel et al. (2016b)
Ariel et al.	2015	West Yorkshire BWV	West Yorkshire Police	Body Worn Video	Y	Article published - Ariel et al. (2016b)

Ariel et al.	2015	PSNI BWV RCT	Police Service of Northern Ireland	Body Worn Video	Y	Article published - Ariel et al. (2016b)
Ariel and Young	2015	BWV Trial	Ventura, USA	Body Worn Video	Y	Article published - Ariel et al. (2016b)
Matheson, J et al.	2015	Project 360 - Domestic Violence second responder RCT	Leicestershire Police	Domestic violence	N	Initial report published - full results awaited
Goosey, J. et al.	2013 to 2015	Bracknell Forest DV RCT	Thames Valley Police and Bracknell Forest Council	Domestic violence	Y	Report complete as M.St. Thesis: analysis showed reduction in Crime harm for Treatment group
Ariel, B., Weisburd, D. and others	2015	Tel Aviv Hotspots Experiment	Tel Aviv, Israel	Hotspots	N	Complete - report awaited
Groff, E.	2015	Offender Contact in violent Hot Spots	Philadelphia, USA	Hotspots	N	Field work complete - reported at Cambridge EBP Conference 2015
Ratcliffe, J.	2015	Predictive Policing	Philadelphia, USA	Predictive Policing	N	reported at Cambridge EBP Conference 2015
Miller, J. and Alexander, B	2015	PJ Stop and Search training	National College of Policing, UK	Legitimacy	Y	Field work complete - reported as Giacomontonio et al. (2016).
Santos, R.B. and Santos, R.G.	2015	Offender-focused police intervention in residential burglary and theft from vehicle hot spots	Port St Lucie, Florida	Hotspots	Y	Complete and published - Santos and Santos (2016)
Weisburd, D.W and Gill, C.	2015	Smart Policing Collective Efficacy RCT	Brooklyn Park, USA	Hotspots	N	In progress - personal communication - analysis phase until first quarter 2017

Sue Ming-Yang and Gill, C.	2015	Improving Police Response to Mental Health Crisis in a Rural Area	Roanoke, USA	Mental Health	N	In progress - personal communication - field work 2016-17
Ruda, S., Behavioural Insights Team (BIT) and Avon and Somerset Police	2015	Police Recruiting Experiment	Behavorial Insights Team and Avon and Somerset Police	Police Recruiting	Y	Complete pending report publication
Routledge, G., Barnes, G. and Durham Police project team	2016	Operation Checkpoint	Durham Police	Pre-court diversion	Y	Field work in progress
Jaitman, L.	2015	Montevideo hotspots experiment	IDB	Hotspots	N	in progress - reported at ASC 2015
Jaitman, L.	2015	Montevideo hotspots experiment - tactics response test	IDB	Hotspots	N	in progress - reported at ASC 2015
Uchida et al.	2015	Los Angeles Body Worn Video experiment	LAPD and UCLA/GMU	Body Worn Video	N	in progress - reported at ASC 2015
White et al.	2015	Spokane and Tempe BWV experiment	Arizona State University	Body Worn Video	N	Published: White,M.D., Gaub, J.E and Todak, N. (2017) Exploring the potential for Body Worn Cameras to reduce violence in Police-Citizen Encounters, *Policing,Online First, 1-11.*
Braga et al.	2015	Las Vegas BWV experiment	Harvard/Rutgers/CAN Institute for Public Policy Research	Body Worn Video	N	in progress - reported at ASC 2015
Denley, J. and Ariel, B.	2015	West Midlands Organised Crime Group RCT	West Midlands Police and Cambridge University	Offender Management	Y	Field work in progress

Stanko, B. and Dawson, P.	2015	Nudging Offenders:	Metropolitan Police and Mayor's Office for Policing	Offender Management	Y	Reported at SEBP 2016
Demir, M.	2016	Recorded Justice: A RCT of the Effect of Body Worn Cameras on Police and Citizens	Eskisehir, Turkey (Rutgers University)	Body Worn Video	Y	PhD Thesis at Rutgers - https://rucore.libraries.rutgers.edu/rutgers-lib/50514/ - author was for 20 years a police officer in Turkey
Magnusson, Mia-Maria	2016	Stockholm Drug Offender experiment	Stockholm Police, Sweden	Offender Management	Y	Data gathering complete - article due in 2017
Jorrat, D., Ortega, D. and Ronconi, L.	2015	Argentina Use of Force training experiment	Development Bank of Latin America	Police Training	N	in progress - reported at ASC 2015
Collazas, D., Garcia, E., Meija, D., Ortega, D. and Tobon, S.	2016	Medellin Hotspot Experiment	Universidad de los Andes	Hotspots	N	in progress - reported at ASC 2015
Baker, S. and Fellows, K.	2016	Birmingham NIP Nudge Experiment	Behavorial Insights Team and West Midlands Police	Offender Management	Y	Using Nudge theory to intervene in driver behaviour
Whitehouse, J. and Fellows, K.	2016	Birmingham Witness experiment	Behavorial Insights Team and West Midlands Police	Victims and Witnesses	Y	Using Nudge theory to encourage witnesses to come forward
Swallow, L. and Smallwood, J.	2016	Honesty Prompts Nudge Experiment	Behavorial Insights Team and West Midlands Police	Victims and Witnesses	Y	Nudge theory test of effectiveness of providing honesty promots to victims reporting theft of mobile phone
Murray, A. and Foster, P.	2016	Fresh Start Offender Management RCT	Behavorial Insights Team and West Midlands Police	Offender Management	Y	Replication of Stanko and Dawson Nudge RCT - still in pre-test phase
BIT and WMP	2016	Wrtitng on the Wall: Using Nudge to change offender behaviour	Behavorial Insights Team and West Midlands Police	Offender Management	Y	Field Work Complete and report in draft

Doyle, K	2016	BME Recruitment RCT	Behavorial Insights Team and West Midlands Police	Police Training	Y	Replication of the BIT Avon and Somerset RCT
Henderson, J.	2016	Tread Finder RCT	Metropolitan Police	Forensic Science	Y	Effectiveness of new Footwear matching technology in detecting and preventing crime
Litmanovitz, Y.	2015	Israeli Border Police Training RCT	Israeli Border Police	Police Training	N	Testing two different approaches to training procedural justice to Israeli Border Police - Cluster randomised design
Bedford, L. and Raison, G.	2016	Queensland Police Service Mobility Evaluation	Queensland Police	Technology	Y	Experiment in progress
Arnold, N. and Bedford, L.	2016	Queensland Police Service Body Worn Camera Evaluation	Queensland Police	Body Worn Video	Y	Experiment in progress
Clare, J., Ready, J. and WAP EBP Unit	2016	Body Worn Video experiment	Western Australia	Body Worn Video	Y	Experiment managed by a Senior Sergeant and EBP Unit
Henry, P. and WAP EBP Unit	2016	Operation Communique: Roads Policing Experiment	Western Australia	Roads Policing	Y	Testing the use of targeted warning notices to repeat offending drivers
Andersen, J.P. and Gustafsberg, H.	2016	A Training method to improve Police Use of Force Decision-making	Finnish Police	Police Training	Y	Published on Sage Open
Andersen, J.P. and Gustafsberg, H.	2017	A Training method to improve Police Use of Force Decision-making	Ontario, Canada	Police Training	Y	Replication of the Finnish trial - due to be published in 2017

Chivers, B.	2017	Nudging Up Appearances: randomised control trial encouraging defendants to appear at court via text messages	Hampshire Constabulary	Police Efficiency	Y	M.St. thesis topic - piloting under way in December 2016
Linas et al. (BIT and Avon and Somerset)	2016	Avon and Somersett Body Worn Video Experiment	Avon and Somerset Police	Body Worn Video	Y	Field Work Complete and report in draft
BIT (Australia)	2016	Nudging Up Appearances: randomised control trial encouraging defendants to appear at court via text messages	New South Wales Police, Australia	Police Efficiency	N	Field Work Complete and report in draft
BIT and MPS	2017	Reducing Missing Person Cases	Metropolitan Police	Police Efficiency	Y	Pre-Test preparation
BIT and MPS	2017	Reducing Wanted Offenders and non-appearance	Metropolitan Police	Police Efficiency	Y	Pre-Test preparation
BIT and MPS	2017	Writing on the Wall: Using Nudge to change offender behaviour	Metropolitan Police	Offender Management	Y	Replication of the WMP trial
BIT and MPS	2017	Domestic Abuse reoffending trial	Metropolitan Police	Domestic violence	Y	Pre-Test preparation
BIT and MPS	2017	Reducing Fraud and cyber crime	Metropolitan Police	Cybercrime	Y	Pre-Test preparation
BIT and MPS	2017	Reducing Mobile phone theft	Metropolitan Police	Crime Prevention	Y	Pre-Test preparation
BIT and MPS	2017	Reducing alcohol related violence	Metropolitan Police	Crime Prevention	Y	Pre-Test preparation
BIT and MPS	2017	London NIP Nudge Experiment	Metropolitan Police	Offender Management	Y	Replication of the WMP trial
BIT and MPS	2017	BME Recruitment RCT	Metropolitan Police	Police Recruiting	Y	Three connnected RCTs

Southall, E., Grossmith, L. And Dawson, P.	2016	London MAST trial	Metropolitan Police and Mayor's Office for Policing	Police Training	Y	Test of effectiveness of Multi-agency Mental Health training - published at https://www.london.gov.uk/sites/default/files/mopac_mast_report_march_2016.pdf

APPENDIX 3: Interview protocol for Operation Turning Point:

Project: Operation Turning Point:

Time of Interview:
Date:
Place:
Interviewer:
Interviewee:
Role/Rank of interviewee:

Opening:

Confidentiality

Questions:

1. What is your role?

    How has that role involved you in Operation Turning Point?

2. [Research motivation] why do you think that West Midlands Police have undertaken Operation Turning Point?

    How have you come to that view? [Prompts: training, briefing, meetings, force website, documents]

3. [Experimental station] how do you see Operation Turning Point fitting with West Midlands Police strategy?

    Has that changed as the experiment has progressed?

4. [Researcher role] what do you think are the most important things that the (Cambridge) research team have done to make the experiment work?

    Why do you think that?

5. [Leadership commitment] How far are your managers committed to making Operation Turning Point work?

    [Prompt if not addressed] Why do you say that?

6. [Implementation process] the project has been implemented in Stages over the last 2 years – do you think it is important that the experiment was introduced that way?

[Prompt] Please explain why?

7. [Police culture and professionalism] How far do you think that the experiment has developed professional skills and knowledge?

[Prompts] How? Why?

8. [Police culture and professionalism] Giving police officers professional discretion in the right way has been a key theme of Turning Point: how well do you think that theme has been managed?

9. [Inter-agency decision-making] How important has it been that the police are the sole decision-maker in Operation Turning Point?

10. [Learning from Turning Point] how do you think that the force should learn from Operation Turning Point?

How have you, personally, learnt from it?

Closing:

Are there are other things about Turning Point that you feel we should have covered?

Thank you for participating in this interview. I would like to remind you that your responses are confidential. When all the interviews have been completed, the research team will be transcribing them and analyzing them in order to build a model of how to conduct and manage experiments in policing. Published data will take care, in using the interviews, not to identify individuals.

APPENDIX 4:
Operation Turning Point: an experiment in "offender desistance policing"

West Midlands Police and Cambridge University

Crim-PORT 1.0:

Criminological Protocol for Operating Randomized Trials

INSTRUCTIONS: Please use this form to enter information directly into the WORD document as the protocol for your registration on the Cambridge Criminology Registry of EXperiments in Policing Strategy and Tactics (REX-POST) or the Registry of EXperiments in Correctional Strategy and Tactics (REX-COST).

CONTENTS:

1. NAME AND HYPOTHESES

1.1 **Name of Experiment:**

Operation Turning Point: a randomized trial of "offender desistance policing" in the West Midlands Police area

1.2 **Principal Investigator:**

1.2.1 (Name) Peter Neyroud

1.2.2 (Employer) University of Cambridge

1.3 **1st Co-Principal Investigator:**

1.3.1 (Name) Professor Lawrence W. Sherman

1.3.2 (Employer) Universities of Cambridge and Maryland

1.3.3 2d Co-Principal Investigator

1.3.4. (Name) Barak Ariel

1.3.5. (Employer) University of Cambridge

1.4 **General Hypothesis**:

Offenders who have not been previously been convicted at court, but whom the police would otherwise charge for prosecution, can be more cost effectively dealt with by police-led offender management than by prosecution, subject to a condition of the certainty of prosecution in the event of reoffending or breaking an agreed "contract" about their conduct.

.

1.5 Specific Hypotheses:

1.5.1 List all variations of treatment delivery to be tested:

1.5.1.1 All those arrestees randomly selected for treatment will have a rapid (within 72 hours) diagnosis meeting with a police officer, after which the officer will offer the arrestee the option of not being prosecuted upon the arrestee's agreement to a "turning point contract," unless the arrestee then breaches conditions of the contract or reoffends within 4 months (if the offence is one with a statute of limitations restricting prosecution to 6 months) up to a maximum of 6 months. Reoffending or contract breach will automatically trigger prosecution for the original offence as well as any subsequent offences.

1.5.1.2 the contracts will involve a set of tactics including voluntary curfew, voluntary exclusion zones, voluntary drug and alcohol testing/treatment referral, not associating with named individuals or categories of people.

1.5.2 List all variations of outcome measures to be tested:

1.5.2.1 Frequency of reoffending within 12 months/2 years and frequency of reconviction within 12 months/2 years as compared between the treatment and control group.

1.5.2.2 Frequency of compliance with the agreed contracts of the treatment group, including measuring the compliance levels with different contract tactics (as at 1.5.1.2) –

1.5.2.3 the sentences given to the control group and the level of compliance with sentences.

1.5.2.3.1 the level of victim satisfaction comparing those allocated to the treatment and control groups, subject to the availability of funding for this element.

1.5.2.3.2 the costs to the criminal justice agencies of the treatment and control groups.

2. ORGANIZATIONAL FRAMEWORK

2.1 Multi-Agency Partnership: West Midlands Police delivers treatments with an independent research organization (Cambridge University) providing random assignment, data collection and analysis

2.1.1 Name of Operating Agency: West Midlands Police

2.1.2 Name of Research Organization: University of Cambridge (analysis)

3. UNIT OF ANALYSIS

3.1 People: Offenders arrested by the police and considered to have met the criteria for charging.

3.2 Locations: Offenders will be arrested and taken to one of 3 Custody locations and dealt with one of two Offender Management teams. Data will be gathered to enable analysis of any differences of decision-making, process or outcome between the 3 custody suites and 2 offender management teams.

3.2.1. WMP and Cambridge may seek to expand the area of the trial by phases to include the whole of Birmingham and/or other areas, subject to implementation progress, but this will be treated as a separate experiment. WMP and Cambridge may also seek to expand the trial to include domestic violence and hate crime cases, subject to the agreement of the CPS. This will also be treated as separate experiment.

4. ELIGIBILITY CRITERIA

4.1 Criteria Required

4.1.1 Offenders who have been arrested by West Midlands Police within the 2 Divisions (Birmingham South and Birmingham East) within the trial area and who the custody officer decides satisfy the following conditions: there is sufficient evidence to meet the CPS Code evidential test; they are not considered suitable for informal resolution, caution, Penalty Notices for Disorder (PND) or conditional caution; their case meets the CPS Code threshold as being in the public interest to prosecute; they have no prior court convictions for a criminal offence.

4.2 Criteria for Exclusion

4.2.1 Cases will be filtered out where, despite meeting the criteria in 4.1.1. nevertheless fulfill one or more of the following:

1) where the offender has any previous conviction for a criminal offence;

2) where, if found guilty, the sentence the court is likely to impose in this case, for this offender, will be custodial;

3) all drink-driving offences

4) offences involving the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon '*per s*e'

5) where the consent of the DPP or a Law Officer is required to prosecute;

6) that involves a death;

7) connected with terrorism or official secrets;

8) sexual offences involving offenders or victims aged under 18;

9) hate crime according to CPS policies.

10) domestic abuse cases according to CPS policy

4.2.2. Victims will be consulted as early as possible in the process and, if Domestic Violence and Hate Crime are included as a separate experiment within the trial, victims in these cases will be asked for active, informed consent to the treatment being used. If Domestic Violence and Hate Crime are agreed for inclusion this will be treated a separate experiment within the overall trial.

4.2.3 Offenders will not be required to give informed consent to the trial before randomisation. But given that this means that some offenders selected for treatment may decline the treatment, the level of those declining must not exceed 10%. This issue will be tested in the dry run phase and if the level appears likely to exceed 10%, the Project Manager and Principal Investigator will consider a change to an active consent model before randomization.

5. PIPELINE: RECRUITMENT OR EXTRACTION OF CASES

5.1 Where will cases come from?

Cases will be identified by a 2 stage process: stage 1 – a custody sergeant decides that an offender has met both the evidential and public interest test for prosecution AND that they have no previous court convictions AND that they are not excluded by any of the criteria at 4.2.1: Stage 2 they will be randomized to treatment or control.

5.2 Who will obtain them? As 5.1

5.3 How will they be identified? As 5.1

5.4 How will each case be screened for eligibility? As 5.1 and 4.

5.5 Who will register the case identifiers prior to random assignment? West Midlands Police as above at 5.1 in the Cambridge randomizer

5.6 What social relationships must be maintained to keep cases coming?

5.6.1 Offender managers and principal investigators must stay in close contact with custody officers.

5.6.2 There is a steering group with WMP, Cambridge University and Crown Prosecution Service membership to provide oversight and a working group of frontline staff involved. The Steering Group is linked to the Local Criminal Justice Board within West Midlands, which includes the other criminal justice agencies (Probation, Courts, Witness Service and Defence solicitors).

Additionally, because the Monument Trust has provided the funding for the research, there is a national steering group with senior representatives from the Judiciary, CPS, Police, Parole Board and NGO's.

5.6.3 The protocol is to be tested with a two-phase "dry run" and practice for the custody staff and offender managers before live data collection. The first phase, starting on 16th November 2011 will require all offenders with no prior convictions, whom the custody officer is considering for prosecution, to be entered on the Cambridge Randomiser, which will be set to "all prosecute". This will allow Custody staff to get accustomed to the Randomiser and the decision tree for the experiment. In the second phase, the Randomiser will be switched to "all treatment" and all those within the criteria will be referred to the Offender Managers to provide practice with the process of the Turning Point Contract. The full go live will not be switched until the Project Manager and Principal Investigator are satisfied that sufficient volume has been achieved to iron out initial implementation problems.

5.6.4 There will be weekly correspondence between Cambridge University and WMP during the experiment, with summaries of the cases and progress.

5.6.5 Prior to experiment, the offender managers are to be trained by WMP/Cambridge and other key staff, including custody staff briefed.

5.7 Has a Phase I (no-control, "dry-run") test of the pipeline and treatment process been conducted?

5.7.1 A dry run of the protocol will take place in November/December 2011 as 5.6.3. Full go live and data collection will be subject to the decision of the Project Manager and Principal Investigator.

6 TIMING: CASES COME INTO THE EXPERIMENT IN

6.1 A trickle flow process, one case at a time, with an estimated 40 cases per month in total (control and treatment).

7 RANDOM ASSIGNMENT

7.7 How is random assignment sequence to be generated?

7.7.1. Random numbers case-treatment generator program in secure computer (Cambridge Randomizer)

7.8 Who is entitled to issue random assignments of treatments?

7.8.1 Role: Barak Ariel (via Cambridge Randomizer)

7.8.2 Organization: Cambridge University

7.9 How will random assignments be recorded in relation to case registration?

7.9.1. The format of the Randomiser for the Turning Point experiment is shown at Appendix B. This will record the decisions by Custody Officers, coded to location and officers collar number.

7.9.2. Cases allocated to treatment will be recorded on the WMP Corvus database, kept by the WMP Offender Management team. Cases prosecuted will be recorded on the ISIS database managed by the WMP CJ Department.

7.9.3 Location of data entry: WMP

7.9.4 Persons performing data entry: WMP Offender Management and CJ & Custody teams

8 TREATMENT AND COMPARISON ELEMENTS

8.1 Experimental or Primary Treatment

8.1.1 What elements must happen, with dosage level (if measured) indicated.

8.1.1.1 All the subjects allocated to treatment must have a "diagnosis meeting" with a member of the offender management team within 72 hours of arrest (normally within 24 hours but because of a lack of weekend cover some cases may need an appointment up to 72 hours) and must sign a "turning point contract" setting out the actions, including no reoffending, which they have agreed to following on from the "diagnosis meeting". Cases where these two conditions are not applied cannot be considered to have met the conditions of the treatment.

8.1.1.2 All subjects within treatment who breach their "turning point contract" or reoffend within the agree period of the contract (a minimum of 4 months, up to a maximum of 6 months) must be referred for prosecution. There needs to be a high level of fidelity to this condition because "certainty" of prosecution is a key element of the hypothesis for this experiment.

8.1.1.3. All subjects who accept the treatment but then subsequently decide to change their minds within the contract period must be referred for prosecution.

8.1.2 What elements must *not* happen, with dosage level (if measured) indicated.

8.1.2.1 Arrestees should not be told that they were selected for deferral of prosecution by random assignment. But given that this means that some offenders selected for treatment may decline the treatment, the level of those declining must not exceed 10%. This issue

will be tested in the dry run phase and if the level appears likely to exceed 10%, the Project Manager and Principal Investigator will consider a change to an active consent model before randomization.

8.1.2.2 Offenders who have been allocated to treatment must not be allowed to breach their contracts or reoffend without instant referral for prosecution.

8.1.2.3 CPS must not discontinue prosecutions, where an offender subject to treatment is referred for breach of the contract or reoffending. The decision to prosecute is one independently taken by CPS. It is possible, particularly in assault cases, that there will be some discontinuance. The Project Manager and Principal Investigator will monitor the level of discontinuances closely.

8.2 Control or Secondary Comparison Treatment

8.2.1 What elements must not happen, with dosage level (if measured) indicated.

8.2.1.1. Offenders who are allocated to the control must be charged and referred for prosecution.

8.2.1.2 Offenders who are allocated to control should not be told that this allocation was based on random assignment. However, general information about the trial is being provided to defence solicitors.

9 MEASURING AND MANAGING TREATMENTS

9.1 Measuring

9.1.1 How will treatments be measured? By examining the official record in Corvus, which will include any contracts and any record of their being breached.

9.1.2 Who will measure them? Data will be gathered from WMP systems and analysed by the Principal Investigator

9.1.3 How will data be collected? From WMP operational systems (Custody and CORVUS)

9.1.4 How will data be stored? On secure WMP systems and Cambridge data systems.

9.1.5 Will data be audited? By the CJ Department.

9.1.6 If audited, who will do it? As 9.1.5

9.1.7 How will data collection reliability be estimated? Cambridge calculations

9.1.8 Will data collection vary by treatment type? Data for Treatment will be derived from the Corvus system, data for those prosecuted from the ISIS system.

9.2 Managing

9.2.1 Who will see the treatment measurement data? Management at divisional and force level, the Steering and Working Groups.

9.2.2 How often will treatment measures be circulated to key leaders? Monthly

9.2.3 If treatment integrity is challenged, whose responsibility is correction? The Criminal Justice Department at WMP.

10 MEASURING AND MONITORING OUTCOMES

10.1 Measuring

10.1.1 How will outcomes be measured?

(a) Frequency, prevalence, time-to-failure and harm index level of rearrests and reconvictions as compared between the treatment and control group
(b) Costs to the agencies of prosecution (control group) and offender desistance policing (treatment group). Costs for experimental cases will be estimated by a diary of the offender managers.
(c) If funding is available, interviews with victims of arrestees in both treatment groups will be compared on the same kinds of dimensions as in the WMP ASB experiments.

10.1.2 Who will measure them? Corvus, cost and any victim data to be analyzed under direction of all Co-Principal Investigators by second co-PI

10.1.3 How will data be collected? Data transfers from WMP to Principal Investigators

10.1.4 How will data be stored? In Cambridge secured systems (for offending data) and Cambridge secure systems (

10.1.5 Will data be audited? Yes

10.1.6 If audited, who will do it? WMP CJ Department

10.1.7 How will data collection reliability be estimated?

Sampling of the custody records before, during and after the experiment (both treatment and control groups), for expected numbers, cases included and potential cases excluded. A one month set of sample data of potential cases will be drawn for January 2010 and together with the data from the dry run will be used to provide "expected" data to compare to actuals.

10.1.8 Will data collection vary by treatment type? No.

10.2 Monitoring

10.2.1 How often will outcome data be monitored? Monthly by WMP/Cambridge University by an agreed report process

10.2.2 Who will see the outcome monitoring data? WMP/Cambridge University

10.2.3 When will outcome measures be circulated to key leaders? Monthly

10.2.4 If experiment finds early significant differences, what procedure is to be followed?

Regular reports will be tabled at the quarterly Steering Group and monthly working group. Only the Steering Group will have the power to sanction changes to the protocol.

11 ANALYSIS PLAN

11.1 Which outcome measure is considered to be the primary indicator of a difference between experimental treatment and comparison group?

11.1.1 the comparative harm index of rearrests between the two groups over the first 730 days after random assignments.

11.2 Which outcome measure is considered to be the secondary indicator of a difference between experimental treatment and comparison group?

11.2.1. the comparative costs and benefit ratio of the treatment and control groups as measured by 11.1.1.

11.2.2 Cost-benefit in relation to frequency or rearrest.

11.3 What is the minimum sample size to be used to analyze outcomes?

11.3.1 400 cases (200 treatment and 200 control)

11.4 Will all analyses employ an intention-to-treat framework? Yes

We reserve the option to analyse the data using Instrumental Variables analysis, depending on treatment compliance rates.

11.5 What is the threshold below which the percent Treatment-as-Delivered would be so low as to bar any analysis of outcomes? 80%

11.6 Who will do the data analysis? The 2d co-principal investigator

11.7 What statistic will be used to estimate effect size? Cohen's *D*

11.8 What statistic will be used to calculate P values? t-tests and, if the distribution is appropriate, zero-inflated Poisson regression.

11.9 What is the magnitude of effect needed for a p =. 10 difference to have an 80% chance of detection with the projected sample size for the primary outcome measure. **d= 0.4** (see appendix A for power calculations.)

12 DISSEMINATION PLAN

12.1 What is the date by which the project agrees to file its first report on CCR-RCT? (Report of delay, preliminary findings, or final result).

Preliminary findings will be given to stakeholders within 120 days after completion of experiment and its follow up period.

12.2 Does the project agree to file an update every six months from date of first report until date of final report?

12.2.1. Yes.

12.3 Will preliminary and final results be published, in a 250-word abstract, on CCR-RCT as soon as available?

12.3.1. Yes.

12.4 Will CONSORT requirements be met in the final report for the project? (See

http://www.consort-statement.org/)

12.4.1. Yes.

12.5 What organizations will need to approve the final report?

Cambridge University will provide any conclusions or Aggregated Data it intends to disseminate or transmit to WMP, for review, at least 90 days prior to submitting such materials for publication. WMP shall then have 90 (ninety) days to respond, provide comments and suggestions based on the said materials, whereas Cambridge University agrees to take under full consideration, at the very least in the way of including such comments and suggestions in the disseminated reports.

12.6 Do all organizations involved agree that a final report shall be published after a maximum review period of six months from the principal investigator's certification of the report as final?

12.6.1. Yes.

12.7 Does principal investigator agree to post any changes in agreements affecting items 12.1 to 12.6 above?

12.7.1. Yes.

12.8 Does principal investigator agree to file a final report within two years of cessation of experimental operations, no matter what happened to the experiment? (e.g., "random assignment broke down after 3 weeks and the experiment was cancelled" or "only 15 cases were referred in the first 12 months and experiment was suspended").

Yes. Save conditions stipulated in 12.5 above.

Contact point:

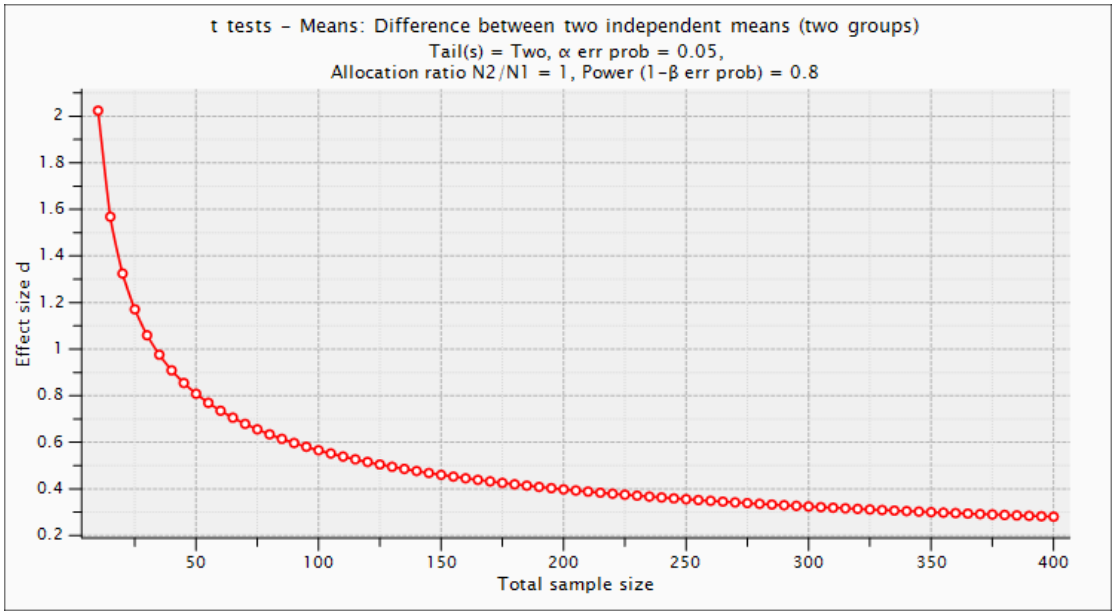
Peter Neyroud CBE QPM,
Institute of Criminology,
University of Cambridge,
Sidgwick Avenue,
Cambridge,
CB3 9DA

Email: pwn22@cam.ac.uk

Appendix A: Power Calculations:

t tests - Means: Difference between two independent means (two groups)

Analysis:	A priori: Compute required sample size	
Input:	Tail(s)	= Two
	Effect size d	= 0.28
	α err prob	= 0.05
	Power (1-β err prob)	= 0.80
	Allocation ratio N2/N1	= 1
Output:	Noncentrality parameter δ	= 2.8139652
	Critical t	= 1.9658827
	Df	= 402
	Sample size group 1	= 202
	Sample size group 2	= 202
	Total sample size	= 404
	Actual power	= 0.8015793



Appendix B: the Turning Point Randomiser

Operation Turning Point Project

Questions

Custody No:

Custody Officers Collar No:

1. Does the offender have any previous conviction for a criminal offence? Yes No

2. Is this offender likely to be sentenced to a period of custody for this/these offences? Yes No

3. Is this an offence of drink/drugs driving? Yes No

4. Does this offence involve the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon *'per se'*? Yes No

5. Is the consent of the DPP or a Law Officer is required to prosecute? Yes No

6. Did this offence contribute to a death of any person? Yes No

7. Is this offence connected with terrorism or official secrets? Yes No

8. Is this a sexual offence involving offenders or victims aged under 18? Yes No

9. Is this offender currently on bail to court for an offence? Yes No

10. Does this offender not have a local address where we are confident they will be staying for the next 4 months? Yes No

11. Does this offence fit the hate crime policy according to CPS? Yes No

12. Does this offence fit the domestic abuse policy according to CPS? Yes No

APPENDIX 5: TREATMENT INTEGRITY STORYBOARD FOR OPERATION TURNING POINT

	STAGE 1: PRE-TEST Nov/Dec 2011	STAGE 2: TREATMENT TEST Jan/April 2012	STAGE 3: 1ST LIVE PHASE May 2012-Feb 2013	STAGE 4: FINAL EVALUATION May 2013-June 2014	COMMENTS
ELIGIBILITY SCREENING	• Screening for low harm 'first offenders' • Screening algorithm embedded in TP Gateway	• CPS review of custody officer compliance • Redraft screening criteria from Custody feedback • Tracking Q.12 exclusions	• Tracking Q.14 exclusions • Tracking completion of the Gateway records against Custody Records	• Instant "kick out" on first negative in Gateway • Tracking Q15 Exclusions	Eligibility screening and Override question evaluated by Hobday (2015)
CASE FLOW	• Pre-test case flow assessment • Inclusions of non-admissions as well as admissions	• Reviewing Gateway cases and exclusions in bi-weekly meetings	• Expansion to Birmingham E&W and Birmingham N • Extending eligibility to "one conviction"	• Inclusion of summons and CPS advice cases	• Domestic Violence and Hate Crime cases excluded – CPS asked for delay on Hate Crime;
RANDOM ASSIGNMENT	• Gateway Mark 1.0 (Figure 2. • Custody staff trained	• Review of Mark 1.0 operation (DOC/TP/) • Revised Gateway Mark 2.0	• Gateway Mark 2.0 – until 17/4/2013 • Review of assignment attrition from Gateway Mark 2.0 • Pause and commitment to restart the experiment with	• Gateway Mark 3.0 - 18/4/2013-23/6/2014 • Block Random Assignment of crime with personal victims and not	• Victim's team and victim script deployed in Stage 4 (DOC/TP/) (Slothower, 2014)

TREATMENT INTEGRITY STORYBOARD FOR OPERATION TURNING POINT

			revised Gateway process – February 2013 (DOC/TP/)		
TRAINING/BRIEFING	• Pre-test training for custody and OMT staff at Cambridge	• Newsletter (DOC/TP/) • Revised training pre-go-live Stage 3	• Training for Birmingham W&C and Birmingham N and partners	• SEBP seminars • Newsletters • Field training and briefing by researchers	Slothower, Sherman and Neyroud (2015) explored model of training/tracking
TREATMENT DESIGN	• Research proposal providing outline (DOC/TP/) • CRIMPORT (Neyroud, 2011) • Project Board agreement to core design	• Testing treatments 'with discretion' • Increased focus on "SMART" treatments	• Treatments supported by 'guidelines' and additional training • Initial assessment of SMART	• "prescribing tool" with embedded treatment details • tracking of "SMART"-ness of treatments by OMT supervisors	Slothower (2015) assessed steps to "SMART" treatment
TREATMENT DELIVERY	• use of the treatment resources available to the OMTs for IoM and to Youth Offending Service for juveniles	• Review of Stage 2 treatment records • Interviews with a sample of offenders and victims	• Treatment guidelines developed • Additional treatments and treatment providers • Decision by OMT supervisors to	• Deployment of the "prescribing tool" for adult treatments • Field testing of LS-CMI tool to support assessment of	

			develop "prescribing tool"	adult offender needs and risks	
TRACKING	• Custody records tracked against Gateway records • Weekly meetings to discuss and debrief Gateway cases and exclusions	• Bi-weekly operational meetings • Field visits from research team • OMT sergeant meetings • Addition of field researcher	• Absence of field researcher (August 2012-January 2013)	• Reinsertion of Field researcher • Master data sheet to track cases and treatments • Treatments assessed for "SMART"	
EVALUATION	• Gateway designed to provide screening and demographic data • Detailed data requirement (DOC/TP/) and data sharing agreement between Cambridge and WMP	• Development of treatment records to provide 'criminal career history' data • In house WMP staff gathering data on shared drive on CORVUS offender management system (for adult offenders)	• Addition of part time Data analyst • Field researcher reviewing data and checking Gateway assignments v. treatments and quality and quantity of treatments • research team taking over prime responsibility for data	• Compilation of the master sheet and data tables	• Victim-offender overlap study (Neyroud, E., 2015)

TREATMENT INTEGRITY STORYBOARD FOR OPERATION TURNING POINT

Appendix 6: The Turning Point Documents

1. DOC/TP/1: Research proposal to West Midlands Police
2. DOC/TP/2: Project Board minutes from 10/5/2011
3. DOC/TP/3: Project Board minutes from 17/6/2011
4. DOC/TP/4: email 7/11/2011 from researcher about post random assignment consent
5. DOC/TP/5: Cambridge Gateway Mark 1.0
6. DOC/TP/6: email 11/7/2012 from field researcher about issues requiring action
7. DOC/TP/7: Turning Point Quarterly Newsletter Issue 1 October 2012
8. DOC/TP/8: Victim's script
9. DOC/TP/9: email 14/6/2011 from Project Manager on pre-test case-flow
10. DOC/TP/10: Project Board Minutes from 18/7/2011
11. DOC/TP/11: the Crimport
12. DOC/TP/12: CPS Review report
13. DOC/TP/13: Cambridge Gateway Mark 2.0
14. DOC/TP/14: email 20/11/2011 to researcher about the Turning Point launch
15. DOC/TP/15: Refresher training day agenda May 2012
16. DOC/TP/16: email 26/4/2012 from researcher reporting progress of Stage 2 and go-live for Stage 3.
17. DOC/TP/17: Turning Point Project: Draft Review of Project Systems by field researcher March 2012.
18. DOC/TP/18: Two-week report from field researcher 14/6/2012
19. DOC/TP/19: Cambridge Gateway mark 3.0
20. DOC/TP/20: Operation Turning Point Training PowerPoint August 2012
21. DOC/TP/21: Offender Desistance Policing PowerPoint
22. DOC/TP/22: email 22/8/2012 from the researcher on next steps and progress
23. DOC/TP/23: Turning Point Project Review of Project Systems III February 2013
24. DOC/TP/24: email 8/4/2013 from field researcher on issues with integrity in Stage 3
25. DOC/TP/25: email 14/5/2013 from field researcher reporting progress.